NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

MENTAL MODELS, TRUST, AND RELIANCE: EXPLORING THE EFFECT OF HUMAN PERCEPTIONS ON AUTOMATION USE

by

Andrea M. Cassidy

June 2009

Thesis Advisor:	Lawrence G. Shattuck
Second Reader:	Nita L. Miller

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
June 2009

3. REPORT TYPE AND DATES COVERED
Master's Thesis

4. TITLE AND SUBTITLE Mental Models, Trust, and Reliance:
Exploring the Effect of Human Perceptions on Automation Use

5. FUNDING NUMBERS

6. AUTHOR(S) LT Andrea M. Cassidy

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING /MONITORING AGENCY NAME(S) AND ADDRESS(ES)
N/A

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (maximum 200 words)

Today's military increasingly uses automation to perform or augment the performance of complex tasks. Automated systems that support or even make important decisions require human operators to understand and trust automation in order to rely on it appropriately. This study examined the effect of varying degrees of information about an automated system's reliability on mental model accuracy, trust in, and reliance on automation.

Forty-two participants were divided into three groups based on level of information received about the reliability of a simulated automated target detection aid. One group received little information, one group received accurate information, and one group received inaccurate information about the target detection aid's reliability. Each participant completed a series of 120 tasks in which he or she was required to identify the presence of a threat target and then decide whether to use an automated aid for assistance. Results indicate a significant difference between the groups' trust in and reliance on automation. The experimental group that received little information trusted the automation less but relied on it more. These findings, accompanied by observational data collected regarding the formation of mental models, demonstrate the necessity of continued research in the field of automation trust.

14. SUBJECT TERMS trust, automation, reliance, mental model

15. NUMBER OF PAGES
93

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
Unclassified	Unclassified	Unclassified	UU

THIS PAGE INTENTIONALLY LEFT BLANK

MENTAL MODELS, TRUST, AND RELIANCE: EXPLORING THE EFFECT OF HUMAN PERCEPTIONS ON AUTOMATION USE

Andrea M. Cassidy
Lieutenant, United States Navy
B.A., University of Colorado, 2002

Submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE IN HUMAN SYSTEMS INTEGRATION

from the

NAVAL POSTGRADUATE SCHOOL
June 2009

Author: Andrea M. Cassidy

Approved by: Lawrence G. Shattuck
Thesis Advisor

Nita L. Miller
Second Reader

Robert F. Dell
Chairman, Department of Operations Research

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Today’s military increasingly uses automation to perform or augment the performance of complex tasks. Automated systems that support or even make important decisions require human operators to understand and trust automation in order to rely on it appropriately. This study examined the effect of varying degrees of information about an automated system’s reliability on mental model accuracy, trust in, and reliance on automation.

Forty-two participants were divided into three groups based on level of information received about the reliability of a simulated automated target detection aid. One group received little information, one group received accurate information, and one group received inaccurate information about the target detection aid’s reliability. Each participant completed a series of 120 tasks in which he or she was required to identify the presence of a threat target and then decide whether to use an automated aid for assistance. Results indicate a significant difference between the groups’ trust in and reliance on automation. The experimental group that received little information trusted the automation less but relied on it more. These findings, accompanied by observational data collected regarding the formation of mental models, demonstrate the necessity of continued research in the field of automation trust.

THIS PAGE INTENTIONALLY LEFT BLANK

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THIS PAGE INTENTIONALLY LEFT BLANK

EXECUTIVE SUMMARY

Human-automation teams are becoming increasingly prevalent in the military. From typical automation roles such as supervisory control, to more recent developments such as unmanned vehicle operation, humans and automation work together on a daily basis. With growing cooperation between humans and machines and increasing complexity of automation comes expanding variability in performance of human-machine teams.

Augmentation of humans with automation influences the design of weapons systems and platforms. Systems are being designed for operation by a smaller crew, and with different training requirements for operators. However, even the best automation can be unreliable and untrustworthy at times. As a result, loss of trust in an automated system acting as part of a human-machine team may negatively affect the team's overall performance, just as when a human partner proves untrustworthy. Although automated aids are becoming increasingly reliable, they are far from perfect. The most advanced automation still requires humans to identify and interpret failures.

This study evaluated three related aspects of human-machine team performance by testing for differences in mental models of, trust in, and reliance on automation between groups possessing different levels of information about an automated aid. The experiment, incorporating a between-subjects design, was conducted in the Human Systems Integration Lab at the Naval Postgraduate School. Forty-two participants were divided into three groups based on the level of information they received about the reliability of a simulated automated target detection aid. One group received little information, one group received accurate information, and one group received inaccurate information about the target detection aid's reliability. In reality, the automated aid's reliability was the same for every group. Each participant completed a series of 120 tasks in which he or she was required to identify the presence of a threat in an image taken from an unmanned aerial vehicle. The experiment occurred in three phases of 40 target detection tasks each. Following the presentation of each image, participants decided whether to use an automated aid for assistance with target detection.

Statistical analyses yielded a significant difference between the groups’ trust in and reliance on automation. The experimental group that received little information trusted the automation less but relied on it more. This result, while surprising, illustrates the complexity of the human-machine relationship and suggests the need to train operators in the appropriate use of automation. These findings, accompanied by observational data collected regarding the formation of mental models, demonstrate the necessity of continued research in the field of automation trust.

The results of this study show we have much to understand about the interrelationships among mental models, trust, and reliance between humans and automated systems. The complicated nature of each of these interrelated features requires a broader and deeper understanding in order to design, build, and operate effective human-machine systems.

ACKNOWLEDGMENTS

First, I thank my thesis advisor, Dr. Lawrence Shattuck, for his encouragement and guidance throughout my thesis development. His positive advice, tempered by Dr. Nita Miller’s critical insights, was critical to my success. Additionally, Dr. Ji-Hyun Yang, Mr. Jeffrey Thomas, and Ms. Diana Kim in the Human Systems Integration department were helpful in shaping my experiment design and executing the experiment. LtCol Anthony Tvaryanas, himself a busy PhD student, also assisted with my data analysis. I am grateful to each of these people for sharing their valuable time for my benefit.

Next, I thank many of the Operations Research department faculty for their support of and patience with me. Dr. Quinn Kennedy, Dr. Susan Sanchez, Dr. Robert Koyak, Dr. Lyn Whitaker, and Dr. Samuel Buttrey each helped me understand some aspect of my data. Dr. Ronald Fricker was so generous with his time that he even responded to my questions via email while on vacation in Hawaii. The talent brought to bear on behalf of a student clearly in over her statistically challenged head inspires me.

Finally, I thank my husband Brian and daughter Claire. Their support, unfettered by understanding, of my work was the biggest contributor to my success as a student here. They make home a happy place where I always feel appreciated and most importantly, loved.

THIS PAGE INTENTIONALLY LEFT BLANK

I. INTRODUCTION

A. PROBLEM STATEMENT

Human-automation teams populate the future of the United States Navy. From standard arrangements such as propulsion control to more recent developments such as unmanned vehicle operation, humans and automation work together on a daily basis. With increasing cooperation between humans and machines and increasing complexity of automation comes increasing variability in performance of human-machine teams. Automation now assists in several areas of task performance, from initial information acquisition to analysis of options, to selecting and implementing a course of action (Sheridan, 2002). This augmentation of humans with automation means we can design weapons systems and platforms with a smaller crew in mind, or with different training requirements for operators. However, even the best automation can be unreliable and untrustworthy at times. Consequently, loss of trust in an automated system acting as part of a human-machine team may have detrimental effects on the team's overall performance similar to those seen when a human partner proves untrustworthy.

Although automated aids are becoming increasingly reliable, they are far from perfect. The most advanced automation still requires humans to identify and interpret failures. The May 2007 grounding of USS ARLEIGH BURKE (DDG-51) provides an example of this continued necessity. During an inbound transit of the Hampton Roads, the destroyer's navigation equipment was functioning properly but was in the wrong mode of operation, giving inaccurate readings for depth and course. Unbeknownst to the crew, they were experiencing a "mode error" (Sarter, Woods, & Billings, 1997), in which no one recognized that the navigation equipment was on the wrong setting for the situation. Despite taking visual bearings that indicated their true position was perilously off the intended track, the ship's navigation team continued to rely on information provided by the automated navigation equipment that indicated they were on course (Fahey, 2007). One reason the navigation team failed to identify the problem was that they did not recognize the automation disagreed with what their eyes were seeing; they trusted that the automation was working properly, which was a correct assessment, but they improperly trusted it when their own senses provided contradictory information. An appreciation for weaknesses inherent in automation is necessary for anyone who works with automation on a regular basis.

Whether automation performs as a decision aid or as a control system, the human involved with it requires some degree of understanding regarding how the automation works. This understanding takes place at different levels depending on the relationship between the human and the automation. For instance, while one can drive a car without knowing how to repair it, a mechanic needs to be able to do both. Similarly, a student may rely on his or her personal computer, but does not need to know how to program it. On the other hand, an Unmanned Aerial Vehicle (UAV) operator needs to understand both how to operate the aircraft as well as what the situation is on the ground below the UAV, potentially thousands of miles away. Various levels of human-machine interaction and of automation autonomy (Sheridan, 2002) combine to form a complex relationship between humans and automation that defies easy categorization. A critical need exists to determine what level of understanding the job requires, and how best to impart that understanding to the human.

A substantial amount of research exists regarding trust in automation, starting with seminal work exploring how human-automation trust compares to interpersonal trust (Lee & Moray, 1992; Muir, 1987; Muir, 1994; Muir & Moray, 1996). Many researchers agree that human trust in automation is an important field of study, but little agreement exists on what constitutes the “right” amount or kind of trust. Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) examined whether human understanding of why an automated decision aid might err contributes to reliance on it, discovering that when human operators knew why an aid could make a mistake, their trust in and reliance on the aid increased, even when the aid’s performance did not warrant the increased trust. Other researchers have investigated trust in automated decision aids relative to human decision-making assistants (Lewandowksy, Mundy, & Tan, 2000; Madhavan & Wiegmann, 2007). Still other researchers have explored how an operator’s mental model of a task affects his reliance on a decision aid (Wilkison, Fisk, & Rogers, 2007; Wilkison, 2008). However, none of this research has examined the relationship between an operator’s mental model of how an automated aid works and the appropriateness of their reliance on the aid. This is an important consideration, as USS ARLEIGH BURKE may never have run aground had the navigator held an accurate mental model that included both the conditions under which the automation's performance degraded, and the symptoms of degradation.

The current study investigated the impact on human-machine team performance of a human operator's mental model regarding an automated aid. First, we discuss the background literature on which contemporary trust in automation research builds. The collection of literature in this effort draws primarily on sources from cognitive psychology and human factors domains. Drawing on early theoretical studies, recent research focuses on manipulating human operators' trust in and reliance on automation of various types. Additionally, the current study explores mental models, drawing on disparate areas of study to form an integrated concept.

B. OBJECTIVES

This research examined how performance of the human-machine team is affected by a human operator's mental model of how the automation functions. It also generalizes the findings to future studies of humans and automation. Specifically, this study:

- Assessed the accuracy of a human operator's mental model regarding automation
- Analyzed the effect of varying mental model accuracy levels on trust in automation
- Evaluated the effect of varying mental model accuracy levels on reliance on automation

C. RESEARCH QUESTIONS

- How should we measure mental models of automation?
- Is an accurate mental model of automation associated with trust in automation?
- Is an accurate mental model of the automation associated with increased use of automation?

D. HUMAN SYSTEMS INTEGRATION (HSI)

HSI is a field of study and a discipline that has gained U.S. Defense Department attention in recent years because it promises to reduce costs and increase performance. Navy HSI practitioners attempt to reduce costs and increase performance by influencing the design of systems across several domains. In many cases, cost reductions can result when a design includes automation to take the place of human operators. However, automation's impact on performance is slightly more complex. HSI incorporates the study of multiple domains to assess the complex relationship between humans and the systems (automated and otherwise) that they operate, maintain, and supervise. According to the Naval Postgraduate School (NPS) HSI program's website, the domains of HSI are: Human Factors Engineering, System Safety, Health Hazards, Personnel Survivability, Manpower, Personnel, Training, and Habitability (Naval Postgraduate School, 2009). Of these eight domains, five are particularly relevant to the current study.

Human Factors Engineering (HFE) is the primary HSI domain relevant to research on human-machine interaction. HFE comprises a number of specialties, including anthropometry, cognition, and human performance. The current research explores the connection between cognition and performance, as it investigates how an individual's mental model affects the performance of a human-machine system.

Systems Safety is an HSI domain pertinent to human-machine interaction because poor performance in a human-machine team can lead to devastating errors, such as the crew of USS VINCENNES and the people onboard the civilian airliner they mistakenly shot down. The current study examines the contribution of an individual's mental model to improving human-system performance. One assumption of this research is that an improvement in human-system performance helps to reduce errors and increase safety, and as such, is worthy of pursuing.

The Training domain of HSI is germane to human-machine team research because appropriate training can contribute to improved performance. The present investigation has implications for training regarding mental model acquisition and development related to automated information analysis and decision support tools.

The current study is relevant to the Personnel domain of HSI because technology that is more sophisticated may lead to changing the needs for people. Future recruits will have to understand the automated systems they will be required to operate and maintain. Ten years ago, a Center for Naval Analyses report forecast what the Navy is experiencing today:

> [W]e see a growing requirement for a future sailor who is a skilled technician. […] an increasing proportion of the Navy's enlisted force will be sailors whose job descriptions include the following:
>
> - Apply general principles in technical fields
> - Define problems, establish facts, and make decisions
> - Communicate technical problems and solutions (Koopman & Golding, 1999, p. 3)

Today's sailors, such as those stationed aboard Ticonderoga-class cruisers, face the challenge of working with increasingly maintenance-intensive aging technology and increasingly complex developing technology in a single platform.

The final domain relevant to the present study is the Manpower domain. As increased manpower costs are driving the military to seek decreased manning, the solution is sometimes accomplished through increased use of automation. This solution itself can have unintended consequences which shift manpower needs. This shift in manpower manifests itself as a decreased need for operators but an increased need for maintainers. Additionally, personnel working with complex technology will need more and varied training to perform their jobs (Congressional Budget Office, 2007; National Research Council, 2008). The resulting, highly skilled sailors may require less supervision than sailors did 20 years ago (Moore, Hattiangadi, Sicilia, & Gasch, 2002, p. 3), which implies another shift in Manpower needs.

Ultimately, an understanding of how and why humans trust automation is vital to Navy force planning decisions. The HSI discipline is a multi-faceted approach to achieving a more thorough understanding of the relationship between humans, automation, and the Navy.

E. THESIS ORGANIZATION

Chapter II reviews literature on trust in automation, situation awareness, and the acquisition of mental models. Chapter III describes the research methodology and experiment used to test the research questions above. Chapters IV and V present results and analysis, concluding with a discussion of directions for future research regarding trust in automation. Appendix A contains a coding key used for images in the experiment, Appendix B is the demographic questionnaire administered to participants, and Appendix C contains the reliability reference cards used in the experiment. Appendix D provides an example of the reliability assessment worksheets each participant completed during the experiment. Appendix E is the trust questionnaire, and Appendix F shows the debriefing form used at completion of the experiment.

II. LITERATURE REVIEW

A. OVERVIEW

The current study proceeds from a wealth of literature regarding trust in automation. We first review relevant terms in the field of trust, pointing out critical definitions and previous studies that laid the groundwork for the current investigation. We also identify challenges of measuring trust and mental models. Next, we discuss the relationship of mental model to situation awareness. Finally, we present the implications and gaps in the literature that led to the formation of the current study.

B. EXPLORING THE TERMS

1. Trust and Reliance

Confusion abounds in the literature pertaining to trust in, and reliance on, automation. A good illustration of this confusion is found in the index to a collection of work on humans and automation. Under the entry, “Trust” is written, “see also Operator reliance” (Parasuraman & Mouloua, 1996). In attempting to provide measureable concepts for research, many studies have tried to clarify the difference between trust and reliance. In their overview of trust in automation, Lee and See assert, “[T]rust is an attitude and reliance is a behavior” (2004, p. 53). If that is true, then reliance is quantifiable because we can observe behaviors, but trust is not quantifiable because we cannot observe attitudes. Thus, research should focus on objective measures of reliance and subjective measures of trust.

Sheridan notes that trust “can be both an effect and a cause” (Sheridan, 2002). In human-automation terms, repeated use of a system may have the effect of increasing the operator’s trust. Additionally, that trust may cause further reliance by the human on the automation. Additionally, Sheridan points out that when trust is viewed as an effect, it can mean “understandability,” or the ability of an operator to form a mental model of how the automation works. Thus, development of a measure of trust is important in order to design automated systems that encourage appropriate use by humans. This sense of trust, as an issue affecting automation use and affected by the system's design, is how the current study addresses the concept.

2. Levels of Automation

The distinction between various types or levels of automation is an important one to consider in research. Using our earlier example of a UAV operator compared with a college student, the UAV represents a much higher level of automation than does a statistical analysis software package on one's personal computer. Sheridan (2002) presents the following scale of degrees of automation:

Table 1. Degrees of Automation (After Sheridan, 2002, p. 62)

A Scale of Degrees of Automation	
1.	The computer offers no assistance; the human must do it all.
2.	The computer suggests alternative ways to do the task.
3.	The computer selects one way to do the task AND
4.	…executes that suggestion if the human approves, OR
5.	…allows the human a restricted time to veto before automatic execution, OR
6.	…executes automatically, then necessarily informs the human, OR
7.	…executes automatically, and then informs the human only if asked.
8.	The computer selects the method, executes the task, and ignores the human.

We can draw a naval analogy for the varying degrees of automation using shipboard systems. The Aegis Weapons System, for example, can perform several functions at different degrees of automaticity (from Number 1 through Number 6 in the table above) depending on the situation and the commanding officer's direction. On the other hand, the automation involved in controlling the propulsion plant requires more participation from its human monitors.

3. Trust in Automation

Teamwork research acknowledges trust between teammates is critical to effective performance. Trust in teams is simple to gauge, at least subjectively: just ask the teammates. However, measuring trust and its impact on teamwork is more complicated when one of the teammates is a machine. Muir (1987) explores literature regarding trust between humans and relates it to human-machine interaction. She concludes that Barber's (1983, cited in Muir, 1987) explanation of how humans trust other humans also applies to human-machine trust. Muir combines Barber's ideas with those of Rempel, Holmes, and Zanna's (1985, cited in Muir, 1987) to create a hybrid definition of human-machine trust. Barber says interpersonal trust is based on expectations that 1) natural and moral laws persist; 2) those around us are technically competent; and 3) those around us will behave responsibly (fiduciary responsibility).

Rempel, Holmes, and Zanna take a more dynamic view of the nature of interpersonal trust, noting that trust develops over time from predictability to dependability to faith. Based on work by Barber and Rempel et al., Muir (1987) proposes a model for human-machine trust: trust is the expectation held by a member of a system of the persistence of the natural and moral social orders, and of technically competent performance, and of fiduciary responsibility from another member of the system and is related to objective measures of these qualities. However, Muir also points out that trust is based on "the *perceived* qualities of another and is therefore subject to all the vagaries of individual interpretation" (1987, p. 531). Thus, we can gauge an operator's expectations about an automated system through subjective measures while more objectively measuring the system's "qualities" upon which the operator bases his judgments and expectations.

Some research has investigated the relationship between trust and perceptions of reliability as they affect the decision to use automation. For example, Dzindolet, Pierce, Beck, Dawe, and Anderson (2001) equate an operator's perception of an automated aid's reliability to the operator's trust in the aid. They claim that a cognitive comparison of one's own versus an aid's reliability leads to a perception of the automation's utility, such that the operator will consider automation more useful if the operator believes he or she is not as reliable as the aid. This perceived utility in turn leads to “relative trust” in the automation and then to automation use (Dzindolet et al., 2001). Further exploration of these concepts in Dzindolet, Pierce, Beck, and Dawe (2002) produced data that suggested operator trust is more likely to suffer than perceived utility because of a lack of understanding regarding how the automation works.

4. Trust Calibration

Muir (1994) defines the calibration process as setting one’s trust equal to an objective measure of the machine’s trustworthiness. Recognizing that human failures to trust machines appropriately will lead to poor system performance, Muir (1987) recommends increasing a human operator’s trust calibration in several ways, one of which is “improving the perception of trustworthiness.” Lee and Moray (1992) and Muir and Moray (1996) explored this theory by evaluating human operators’ trust in a simulated pump mechanism after it malfunctioned. One notable discovery was that after an initial drop in trust, operators’ trust increased as they became accustomed to the presence of a constant error, indicating that perhaps the operators had calibrated their trust in the automation. In both studies, the objective measure of the pump’s trustworthiness was its performance, which was poor and negatively correlated with trust in the machine. However, as the operators recognized the error was constant and they could compensate for the decreased performance by adjusting their own performance, the operators’ trust increased. These results suggest that humans can indeed calibrate trust to a level warranted by a machine’s performance. Training operators to more accurately perceive automation is likely to improve appropriate trust in it. Exactly how operators acquire those perceptions, or if they can be altered once acquired, remain topics for investigation.

The concept of relative trust is related to the concept of calibration of trust. In their study, Dzindolet, Beck, Pierce, and Dawe (2001) hypothesized that automation use is the “outcome of a comparison process between the perceived reliability of the automated aid (trust in aid) and the perceived reliability of manual control (trust in self)” (p. 8). Relative trust is the name the authors give to the social process that mediates the cognitive one (see Figure 1). If an operator trusts the automated aid more than himself or herself, and perceives it to be more reliable, the operator will use the automation. In this model, the effect of changing perceptions of reliability is unclear, however, and is explored further in the current study.

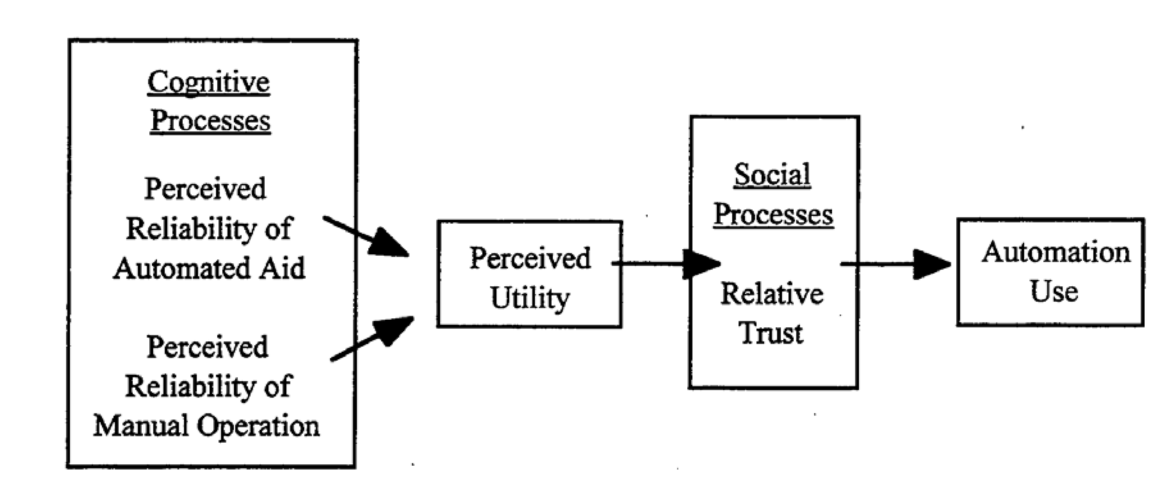


Figure 1. Hypothesized decision process for automation use (After Dzindolet et al., 2001, p. 9)

Expanding on the work of Muir and Moray, McGuirl and Sarter (2006) conducted a study that supports the notion of calibrating trust in automation in order to improve human-machine team performance. In this study, the researchers presented aircraft pilots with either no information or continuously updating information about the status of a decision aid designed to assist with diagnosis of icing on the wings. McGuirl and Sarter found that the pilots with updated information on their decision aid’s status performed significantly better than did the pilots without the status information. These findings suggest trust calibration, or being able to match one’s trust to the capability of decision automation, has an effect on human-machine team performance.

A human operator’s trust in automation is important for proper use of automation and the resultant performance of the human-machine system. However, many researchers acknowledge trust is only one factor in predicting an individual’s appropriate use of automation (Dzindolet et al., 2001; Lee & Moray, 1992; Parasuraman & Riley, 1997). Other factors include preconceived beliefs about automation (believing “only a human could do that task"), mental workload and cognitive overhead (deciding that automation use will actually reduce one's workload instead of increasing it), and self-confidence (thinking oneself capable of performing a task) (1997).

In their discussion of these and other issues affecting human-machine relationships, Parasuraman and Riley (1997) define three different ways humans improperly use automation. *Misuse* occurs when operators rely too much on automation, trusting it when it should not be trusted. *Disuse* happens when operators do not rely enough on automation, ignoring signals and alarms they regard as overly sensitive. Automation *abuse* results when designers or managers apply automation incorrectly or without consideration for its effects on human performance. Ultimately, they advocate for "[b]etter operator knowledge of how the automation works" (Parasuraman & Riley, 1997, p. 248) in order to give the best chance for proper human-automation performance.

Training people in how automation functions seems sensible. Such training may reduce what Sheridan (2002) terms the "magical" nature of automation, that "[t]o a naïve user the computer can be simultaneously so wonderful as to seem faultless, and if the computer produces other than what its user expects, that can be attributed to its superior wisdom" (p. 174). That is, a person without an understanding of how automation works may be more likely to over trust it. In order to investigate this idea, Dzindolet, Peterson, Pomranky, Pierce, and Beck (2003) conducted a series of experiments in which people used a decision aid to identify a camouflaged soldier. Their findings indicate that information about why a decision aid might make mistakes increases reliance on the aid, even if reliance is unwarranted. In order to mitigate these effects and encourage appropriate reliance, Dzindolet et al. recommend providing human operators with both training on how the automation works and experience in using it. Riley (1996) also recommends that because training allows people to understand automation states and anticipate future actions, it may provide the operator with a rational basis for decisions to use automation or not.

We now have some indications regarding what might influence an operator to appropriately trust and use automation. We turn next to issues regarding how to make automation more worthy of human trust. Lee and See (2004) distinguish between *trustworthy* and *trustable* automation this way: trustworthy automation is that which functions efficiently and reliably, while trustable automation functions simply and transparently. This definition of trustable does not imply an operator must understand advanced computer algorithms, but does mean an operator should be aware of more than just how to monitor the automation. Among the recommendations for improving automation trustability, Lee and See list "[d]esign for appropriate trust, not greater trust" (2004, p. 74). However, little research has investigated how to determine what appropriate trust is for different types of automation.

Another study examined the results on human performance of automation at two levels of reliability and at four different levels of information processing (Rovira, McGarry, & Parasuraman, 2007). Automation reliability was 60 percent in the low reliability condition and 80 percent in the high reliability condition. The study used Parasuraman, Sheridan, and Wickens' (2000) taxonomy to classify the types of automation used. The 2007 study used one level of information automation and three levels of decision automation. The three levels of decision automation included low, medium, and high decision support, depending on the level of detail in the automation's recommendation to the operator. Participants performed a series of target engagement tasks in each of the eight combinations. Results indicated that reliable automation improves decision times but also increases operator complacency, which creates a large cost for automation failure. Additionally, although the researchers predicted participants' trust would vary with automation reliability, their findings did not support this, suggesting a dichotomy between automation performance and operator perception. The researchers propose that where decision support automation is unreliable, informing the operator of its unreliability and allowing access to the raw data may lessen the consequences for failure. Although allowing operators such access is one possible option, designers will need to factor this in early in the design process. That may not be possible with older systems that are otherwise useable. A better option to enhance human-machine performance in the interim may be to train operators how to cope with the failings of their unreliable automated counterparts.

Another study (Wiegmann, Rich, & Zhang, 2001) that provides evidence for evaluating trust and reliance separately measured human operators' responses to automation at varying levels of reliability. Participants performed a monitoring task in which they could use an automated aid to assist with diagnosing pump failures in a simulated waste processing plant. Three experimental groups performed a series of 200 trials with three different automation reliability levels as starting points. One group started with 100 percent reliable automation, another with 60 percent reliable automation, and the third with 80 percent reliable automation. The first group's automated aid decreased to 80 percent reliability after 100 trials, while the second increased to 80 percent reliability. The third group's automated aid remained at 80 percent reliability throughout the experiment. Participants were told the automated aid's reliability was unknown, but that it would change during the experiment. Results indicate that human operators' reliance on automation is sensitive to the automation's reliability, and that operators tend to underestimate the automation's reliability regardless of its actual performance. One drawback to this study is that it equates an operator's assessment of automation reliability with an operator's trust in the automation, whereas the two may not be equivalent.

Further breaking down the concept of trust into discrete definitions of reliance and compliance, Madhavan and Wiegmann (2007) explored how people responded to a decision "assistant" who was either human or automated. Rather than equating a person's agreement with an automated decision aid with trust in the automation, Madhavan and Wiegmann break the interaction down into two separate processes. They define reliance as an operator's agreement with a decision aid that a signal is absent, while compliance is an operator's agreement with a decision aid that the signal is present. Their study manipulated the assistants' reliability (70 percent and 90 percent) as well as their level of perceived expertise (novice and expert). Results indicated significantly less reliance on or compliance with an expert automated aid (relative to a human aid) when the aid's reliability was 70 percent. These results suggest that people are less likely to put up with low reliability in automated aids.

Although trust is difficult (if not impossible) to quantify, specific instances of trust have been measured. Using a subjective, but empirically determined (Jian, Bisantz, & Drury, 2000), scale to measure trust in automated decision aids, Bisantz and Seong (2001) investigated the effect of source of automation failure on operator trust. Their experiment involved a target identification task that required participants to identify targets as enemy or friendly. Participants could seek the assistance of an automated information aid or an automated decision aid. Participants were separated into three groups by what they knew regarding potential automation failures: in one, participants were told the decision aid was vulnerable to external sabotage, in another, that the decision aid was vulnerable to internal hardware or software problems, and in the third, participants were told nothing about possible failures. At three intervals over the six experiment trials, participants rated their trust in the automated aid using a seven-point scale anchored at “Not at all” and “Extremely” for each of the following statements.

1. The system is deceptive
2. The system behaves in an underhanded manner
3. I am suspicious of the system’s intent, action, or output
4. I am wary of the system
5. The system’s action will have a harmful or injurious outcome
6. I am confident in the system
7. The system provides security
8. The system has integrity
9. The system is dependable
10. The system is reliable
11. I can trust the system
12. I am familiar with the system

The first five questions are negatively framed, while the last seven are positively framed. This distinction allows for testing of different aspects of trust. Responses to the subjective trust questionnaire indicated operator trust declined less in the group who believed the failure source was external to the automated aid. More importantly for the current study, Bisantz and Seong's work validated the use of a trust questionnaire that was sensitive to different aspects of trust and to different automation failure conditions.

Delving deeper into why humans trust automation, sometimes inappropriately and sometimes appropriately, requires a deeper understanding of several factors, one of which is the concept of mental models, or how people conceptualize the world in which they live.

C. MENTAL MODELS

1. Cognitive Psychology versus Human Factors Approaches

Mental models are an ephemeral concept that researchers in several disciplines have studied. Cognitive psychologists, human factors engineers, and computer designers have all investigated mental models as they relate to their specific areas of study, with little consensus. Cognitive psychologists tend to accept the idea that mental models "enable individuals to make inferences and predictions, to understand phenomena, to decide what action to take, and to control its execution, and above all to experience events by proxy" (Johnson-Laird, quoted in Wilson & Rutherford, 1989, p. 621). This definition of mental models seems to fit with how Muir (1987) describes mental models. She discusses them in relation to the persistence of natural physical laws, viewing a mental model as an understanding of physical processes that allow a human to predict future events. However, the concept of mental models has been only tangential to human factors studies of trust in automation. Wilson and Rutherford (1989) criticize the human factors community's lack of a coherent conception of mental model as contrasted with the psychology community's well-accepted one.

Contemporary human factors interest in mental models revolves around team cognition research (Cannon-Bowers, Salas, & Converse, 2001; Rouse, Cannon-Bowers, & Salas, 1992). Many of these studies examine how mental models or shared cognition affects team performance related to complex systems. Rouse et al. (1992) provide a description of mental models (see Figure 2), outlining three main functions as they relate to human-system relations. The descriptive function pertains to a person's knowledge of the system's purpose and physical description. The explaining function involves a person's knowledge of the system's operation and its current state. The prediction function relates to a person's ability to form expectations about the system's future state and operations.

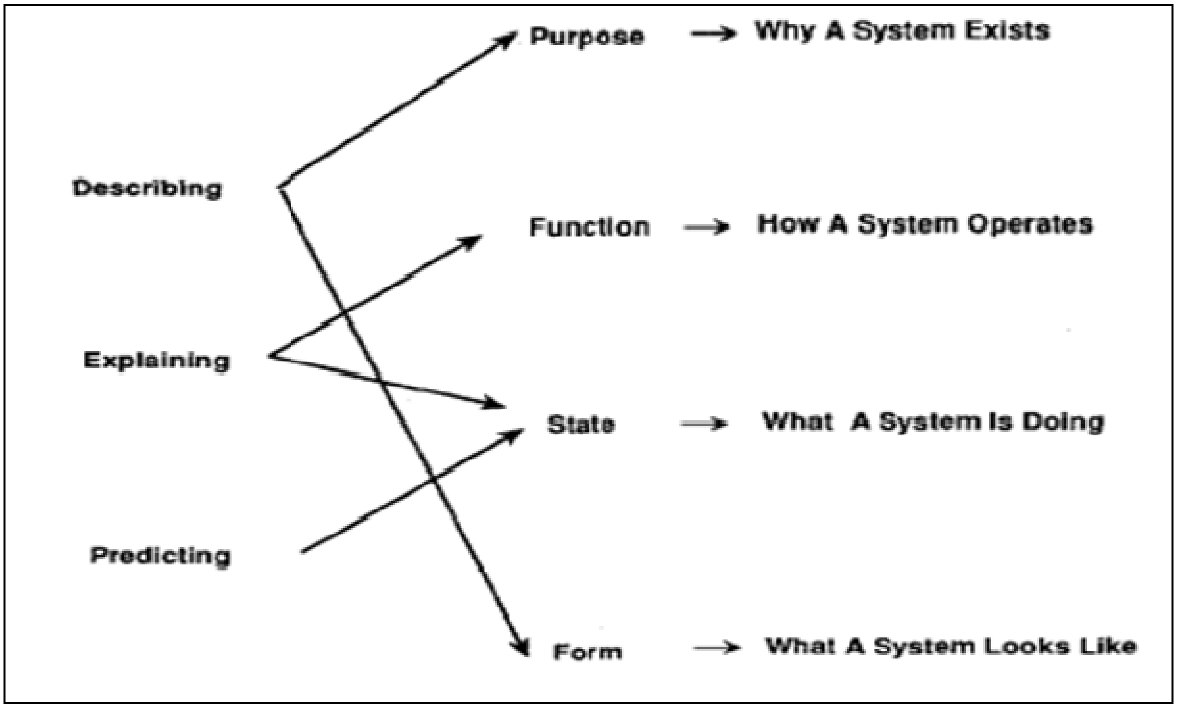

Purpose → Why A System Exists
Describing
Function → How A System Operates
Explaining
State → What A System Is Doing
Predicting
Form → What A System Looks Like

Figure 2. Nature of mental models (From Rouse et al., 1992, p. 1300).

Rouse et al. contend that an understanding of mental models as separate from general knowledge is necessary to understand performance where complex systems (including tasks, equipment, and human teams) are involved. If the descriptive, explanatory, and predictive components of mental models apply to interactions of human teammates, they should apply when a teammate is automated. Additionally, these components may help support an explanation for appropriate human trust in automation when an individual's mental model is properly developed.

One study in particular highlights how the difference between the human factors and cognitive psychology communities' definitions of mental model translates to applied research. Wilkison, Fisk, and Rogers (2007; see also Wilkison, 2008) considered the operator's mental model as central to the issue of trust in automation. Wilkison's study addresses mental model quality at three levels: none, low, and high. However, his definition of mental model is more about understanding the task than the automation's function. Additionally, Wilkison employs a process that builds rather than measures mental model quality. This illustrates the gap between team performance mental model research and psychology-based mental model research. Team mental model research has focused on the sharing among team members of concepts regarding roles and responsibilities (Cannon-Bowers et al., 2001). The cognition-based research Wilkison employs relates to how humans acquire spatial knowledge, and leads to Wilkison's measurement of mental models as levels of "acquisition" of that knowledge. In contrast, the current study uses a concept of mental model more akin to that studied in team research, as an understanding of what roles each team member (in this case, the operator and the automated aid) will fulfill in the execution of a task, and how the performance of one affects the performance of the other.

2. Mental Models versus Situation Awareness

Defining what constitutes a mental model is as elusive as determining how to represent one. Everyone seems to think that such conceptualizations exist in the human mind, but no one seems to know how to represent them or how to use them. Endsley's research on situation awareness (SA) contends that a mental model is general while SA is specific to the circumstances one encounters on a minute-to-minute basis (2000). In her representation, a person's mental model consists of relatively static components that develop with time and experience, while SA is more dynamic and provides input to the mental model, developing it over time (see Figure 3).

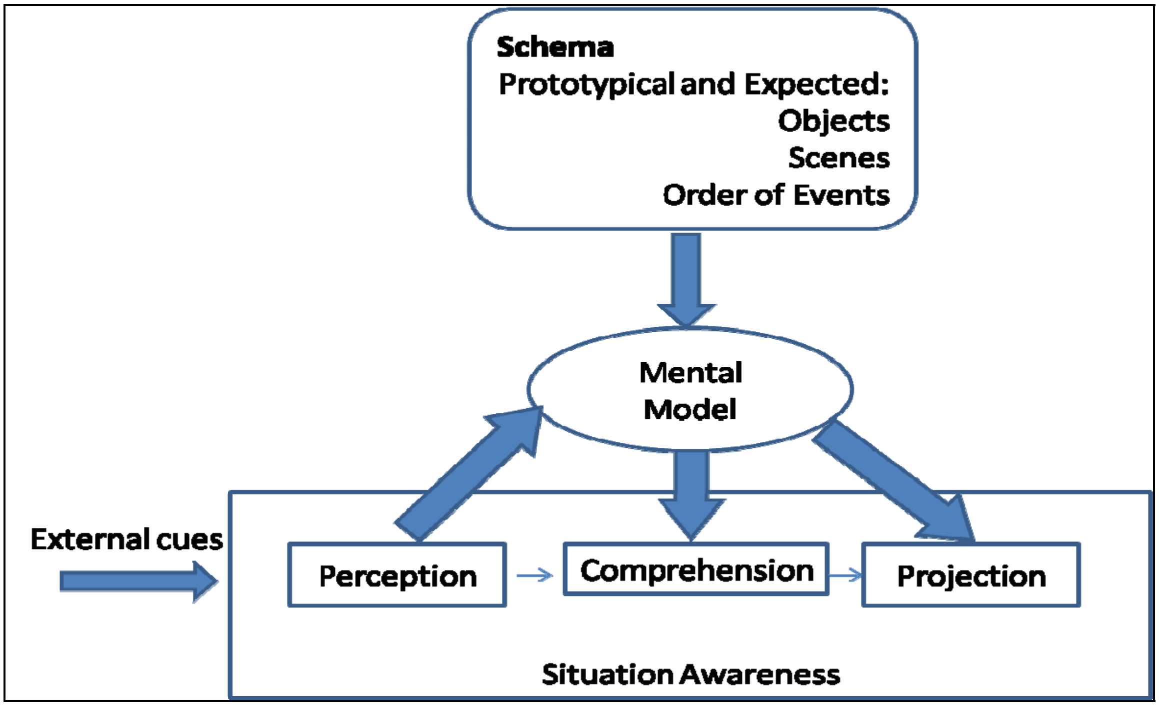


Figure 3. Endsley's model of SA and Mental Model (2000).

If we take Endsley's view on mental models, then we see mental models both affect and are affected by SA. We may improve Figure 3 with a feedback loop indicating the iterative nature of SA's relationship to mental model. If SA is dynamic and changeable, then mental models must have similar qualities. Mental models may not change as frequently as a person's minute-by-minute understanding of a current situation, but a person's mental model will change and develop with experience. That is, one's mental model may not be constantly changing, as is one's SA, but it seems likely that one's mental model is at least adjustable.

Despite Endsley's depiction of mental model as distinct from SA, some research conflates the two ideas, making data collection and analyses difficult. Nunes (2003) set out to examine the effects of new airspace management technology (a predictive aid) on air traffic controllers' mental models. The experiment used the Situation Awareness Global Assessment Technique (SAGAT) to measure SA, which the author points out is used interchangeably with "the current state of the mental model" (p. 66). Additionally, the experiment used response time and accuracy to measure problem solving ability, which the author hypothesizes, will provide insight to the controllers' mental models. The results of the study indicate the predictive aid did not affect the controllers' SA, but also that it may have negatively affected their problem-solving abilities. The author suggests the presence of a predictive aid may actually inhibit a controller's problem-solving skills by reducing the requirement for the controller to develop a mental model of how to accomplish a task. The amount if inference required to reach this interpretation demonstrates the difficulty of measuring mental model as a discrete entity in human factors research.

3. Mental Model as a Factor in Appropriate Trust in Automation

Adjusting one's mental model with experience may lead to trust that is more accurate. Properly calibrated trust in automation should demonstrate itself in the operator's agreeing with automation under the right circumstances, and disagreeing in situations in which automation has shown itself to be unreliable. An operator recognizes a situation he or she is in because he or she has a mental model of both the system and of how the automation behaves in a particular situation. This mental model contributes to a more complete understanding of other components of the environment with which the operator is working. Sheridan (2002) acknowledges the theory is an emerging one, but that an invalid mental model is one possible cause of human error in human-machine systems. The appeal of mental models to explain some variation in human performance stems more from their intuitive logic than from empirical data in support of their existence.

In the current study, mental models are treated carefully, with an acknowledgment that their explanatory powers are limited by their unproven robustness. We define mental model here as the set of rules by which an operator determines when to rely on automation. One's mental model may start out as tenuous, consisting only of information that one receives from a source outside the system. With experience and time, a person adjusts his or her mental model based on accumulated information and interactions. Since the operators in the current study were limited in the time they had to interact with the automation, they could not feasibly develop an accurate mental model without someone pointing out critical external cues to them. As a result, we made the external cues (trees and roads) more salient in an attempt to make up for the limited exposure participants had with the automation.

D. SUMMARY

1. Implications and Gaps

The current study addresses trust as an issue affecting automation use and therefore influencing the system's design. The literature indicates that research into trust in automation should focus on objective measures of reliance and subjective measures of trust in order to design automated systems that encourage appropriate use by humans. We can gauge an operator's expectations about an automated system through subjective measures while more objectively measuring the system's qualities upon which the operator bases his judgments and expectations. The operator's perception of automation's utility in turn leads to relative trust in the automation, and then to automation use (Dzindolet et al., 2001). Dzindolet, Pierce, Beck, and Dawe (2002) suggested operator trust is more likely than perceived utility to suffer as a result of a lack of understanding regarding how the automation works.

Although early work (Lee & Moray, 1992; Muir, 1987; Muir & Moray, 1996) suggested that humans can calibrate trust to a level warranted by a machine's performance, research has yet to indicate the effect of changing perceptions of reliability on trust in automation. Researchers advocate for "[b]etter operator knowledge of how the automation works" (Parasuraman & Riley, 1997, p. 248) in order to enhance human-automation performance. Dzindolet et al. (2003) recommend providing human operators with both training on how automation works and experience in using it. Lee and See (2004) recommend focusing on appropriate trust as opposed to increased trust when designing automated systems. All of this implies that we may be able to improve human-machine performance by telling operators how to interpret automation failures.

We also need a better understanding of mental models, or how people conceptualize the world in which they live. Descriptive, explanatory and predictive components of mental models may help explain appropriate trust in automation when an individual’s mental model is properly developed. The current study uses a concept of mental model similar to that used in team research, as an understanding of what roles each team member will fulfill. Properly calibrated trust should demonstrate itself in the operator’s agreeing with automation when it is at its most reliable and disagreeing when conditions exist to degrade automation’s reliability.

The literature reviewed in this chapter provides a cross-section of the different fields from which this study will progress. The trust in automation literature provides a solid background regarding how people feel about, think about, and behave toward automation. Past research provides coherent definitions of terminology and testable measures of trust that this study will incorporate. Research from multiple disciplines supplies a faceted understanding of the intangible but crucial idea of mental models. This study hypothesizes that there is a difference in mental models between people who have little, accurate, or inaccurate information about an automated device. Additionally, this study investigates the difference in trust between people with different levels of information about an automated aid. Finally, this study tests the hypothesis that there is a difference in reliance on automation between people with varying degrees of information about it.

III. METHODS

A. METHOD OVERVIEW

The experiment consisted of a series of target detection tasks, similar to those used by Dzindolet, et al. (2002). The targets in the current study were white Ford Explorer-type vehicles and personnel wearing black clothing. The experiment used still images captured in Camp Roberts, California, on 22 February 2009 from an Unmanned Aerial Vehicle (UAV). The UAV flight was part of the ongoing Tactical Network Topology experiment that the Naval Postgraduate School conducts quarterly. Figure 4 shows a sample image. The present experiment included 40 different images per phase. Each of the three phases contained the same series of 40 images, presented to participants in random order to minimize a possible learning effect.

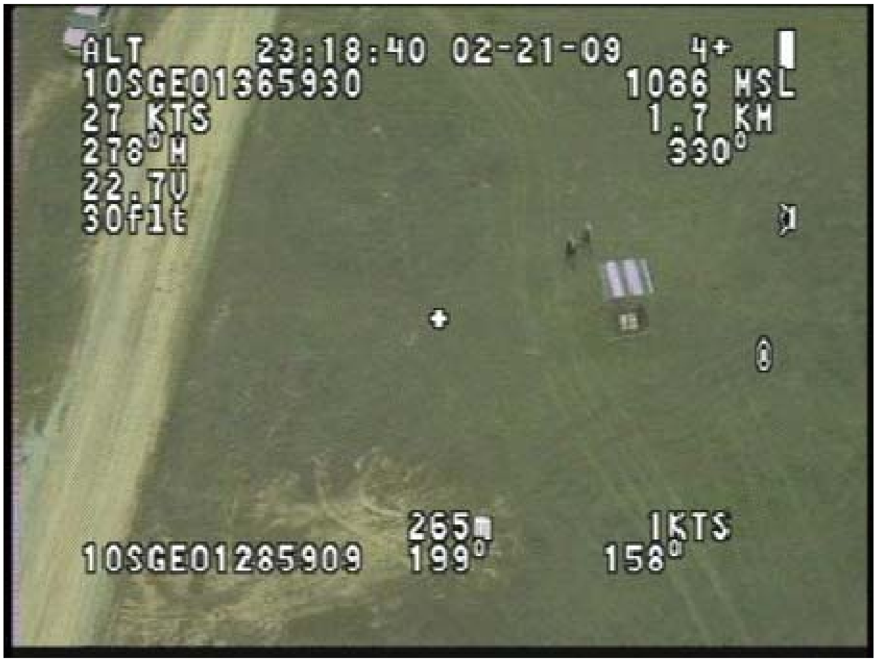


Figure 4. Still image from the experiment.

The experiment employed a target detection device (TDD), which simulated an information acquisition aid (Parasuraman, Sheridan, & Wickens, 2000). To achieve the effect of automation, each of the 40 images was coded for whether or not it contained a) trees and no roads, b) roads and no trees, c) both trees and roads, or d) neither trees nor roads. Prior to the experiment, the researcher determined the images for which the TDD's answer would be correct and incorrect. When images contained neither trees nor roads, the TDD's answer was always correct (100 percent reliable). When images contained only trees or only roads, the TDD's answer was correct nine out of ten times (90 percent reliable). When images contained trees and roads, the TDD's answer was correct eight out of ten times (80 percent reliable). Appendix A contains the coding key for all images used in the experiment. The TDD's reliability in the four conditions remained constant throughout the experiment and across experimental groups while participants' knowledge and ostensibly, mental model, of the TDD's reliability varied between experimental groups.

The researcher told participants the TDD was a limited resource shared with other "analysts." We gave this instruction in order to simulate realistic operational constraints and to prevent participants from relying excessively on the TDD.

This study incorporated a between subjects design with three experimental groups each receiving one different level of the independent variable (mental model accuracy). The dependent variables were trust, mental model accuracy, and reliance on automation. This study evaluated participants' trust in the automated aid using an empirically validated questionnaire developed by Jian, Bisantz, and Drury (2000) to measure human trust in automation (Bisantz & Seong, 2001; Wilkison, 2008). We evaluated mental models by comparing pre-existing information about the reliability of an automated information aid with participants' ratings of the aid's reliability after using it. Participants' reliance on the TDD was collected via E-Prime data logging.

B. PARTICIPANTS

1. Selection

The Naval Postgraduate School Institutional Review Board reviewed and approved the design of this study, satisfying both the Department of the Navy and the American Psychological Association criteria for research involving human subjects. All participants indicated informed consent by signing a form notifying them of their rights as participants in the experiment.

We solicited participants through emails and personal contact. Only active duty U.S. military officers were eligible to participate. The study used a convenience sample from the Naval Postgraduate School population.

2. Demographic Make-up

Thirty-one males and 11 females comprised the participants in this study. Thirty-eight participants were U.S. Navy officers, three were U.S. Army officers, and one was a U.S. Marine. No U.S. Air Force personnel participated. Participants were between 21 and 45 years of age, as shown in Figure 5.

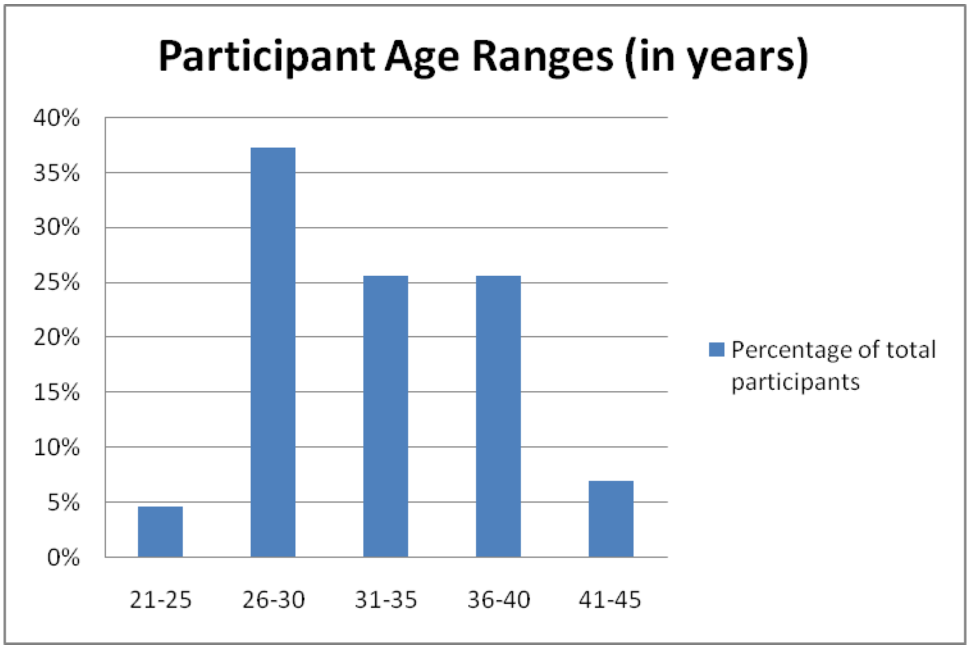


Figure 5. Participant age ranges.

Figure 6 shows participants' military service experience, including enlisted time if the officer was enlisted prior to commissioned service.

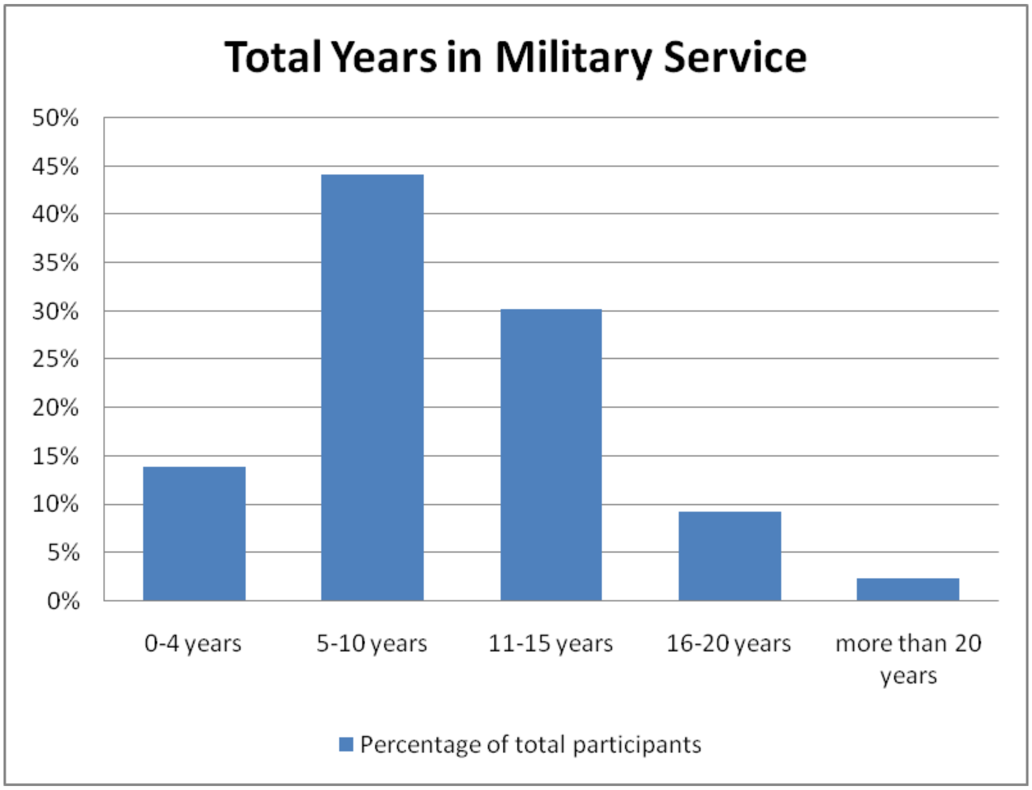


Figure 6. Participants' military service experience.

Most participants had some experience with computer-based games, as indicated in Figure 7. All participants completed the entire experiment. The first participant in Group A was eliminated from the data set because Group A participants' instructions changed following the first experimental run.

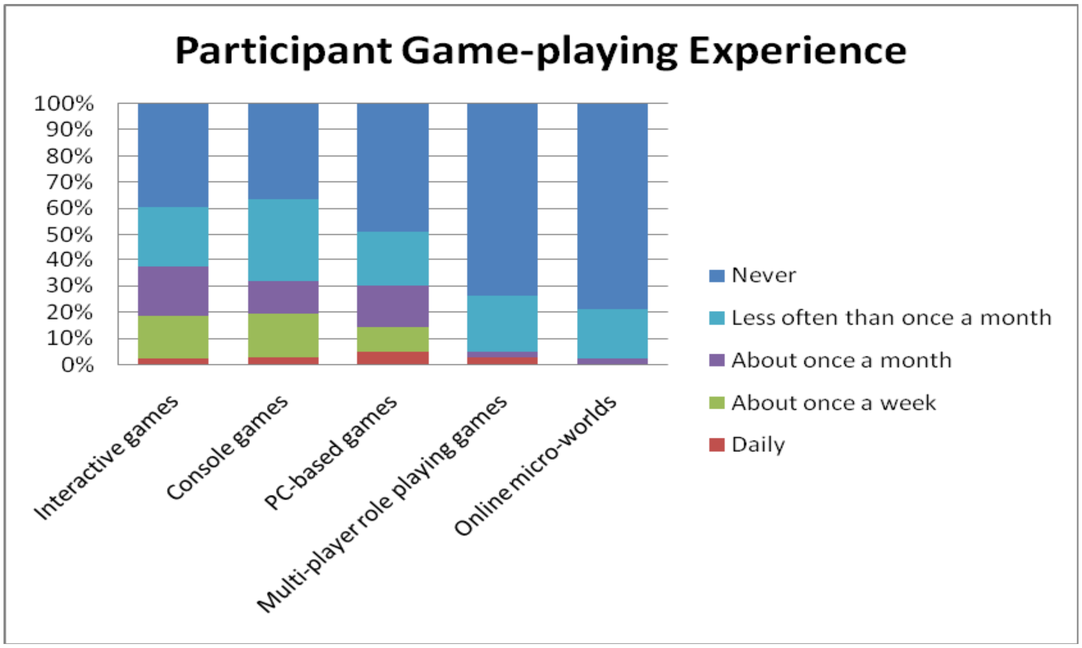


Figure 7. Participants’ game-playing experience.

Additionally, participants indicated how often they used automation in their military jobs before arriving at NPS and now that they are here. Figure 8 shows their responses.

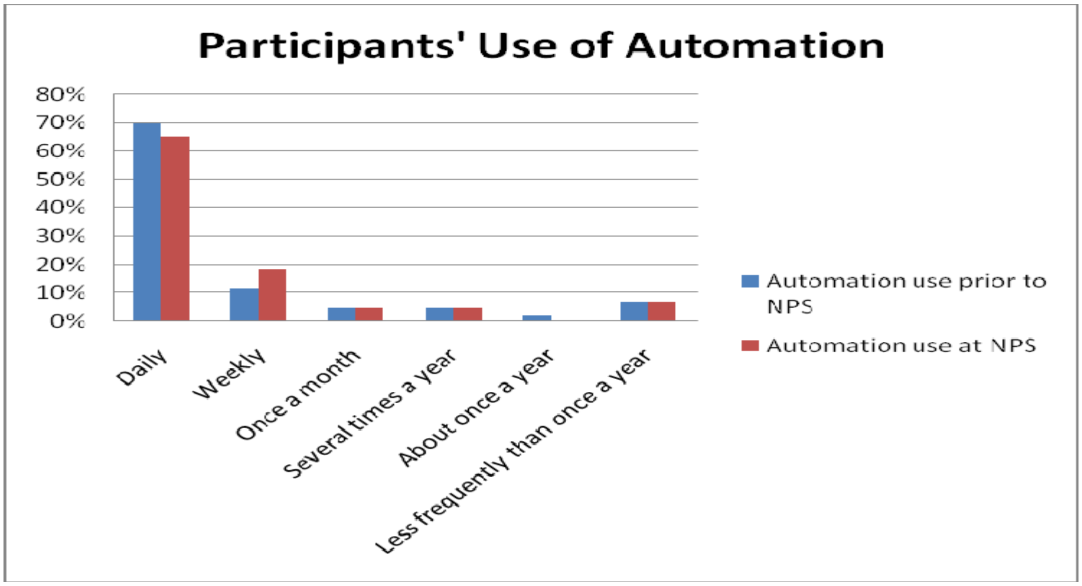


Figure 8. Participants’ automation use prior to and at NPS.

Finally, most participants indicated a high comfort level with using automation, as Figure 9 indicates. Forty-two percent of participants preferred to use automation whenever possible during the military job they held prior to arriving at NPS. Thirty-five percent of participants prefer to use automation whenever possible in their work at NPS.

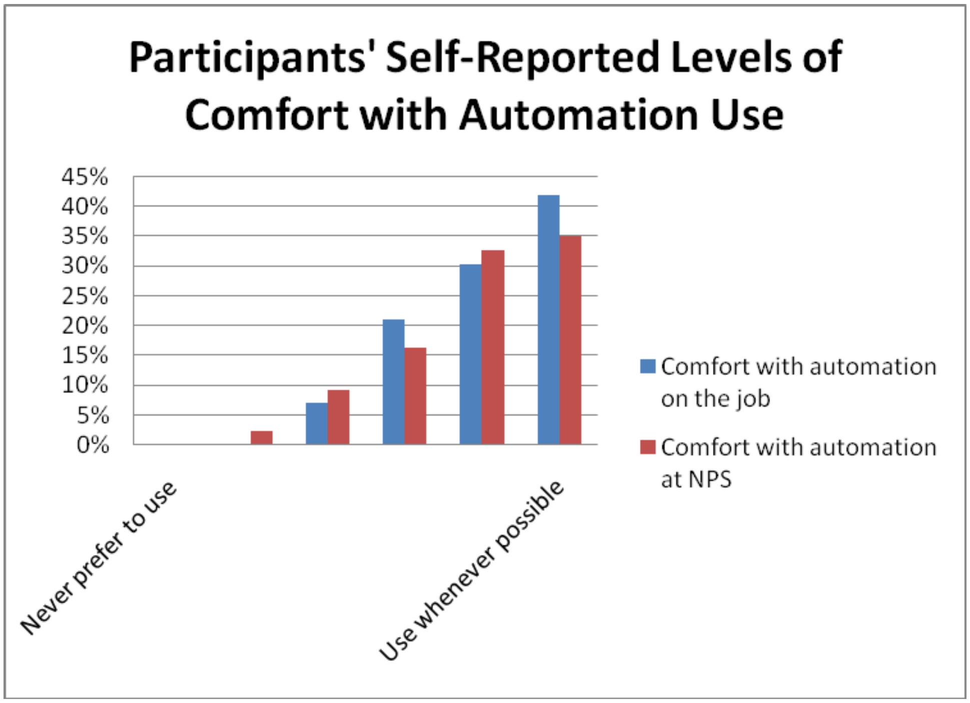


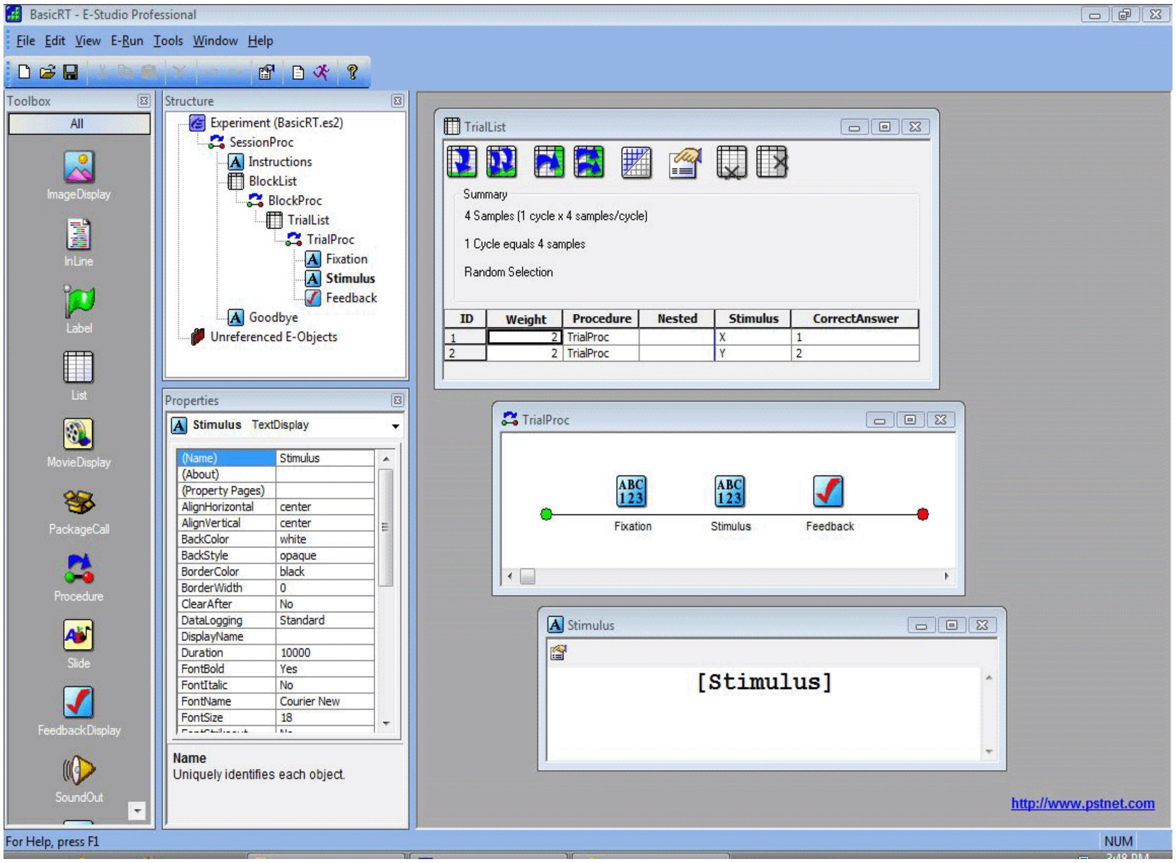
Figure 9. Participants' comfort levels with automation.

C. MATERIALS

1. E-Prime Version 2.0 (Release Candidate)

E-Prime is a set of software applications designed to facilitate conducting research with human subjects. Psychology Software Tools, Inc. first developed the software in 2001 and released the version used in this study in 2007. E-Prime v2.0 (Release Candidate) allows the collection, processing, and analysis of data with the following included applications: E-Studio, E-Basic, E-Run, E-Merge, and E-DataAid (Psychology Software Tools). E-Basic is a programming language similar to Visual Basic. The E-Studio application is a graphic interface that allows a researcher to build and test an experiment, while E-Run runs the experiment and collects the data. E-Merge merges data from multiple sessions and participants into one file for analysis using E-

DataAid or another statistical analysis software package. Figure 10 shows E-Studio's graphic user interface for building an experiment.



Stimulus TextDisplay

(Name)	Stimulus
(About)	
(Property Pages)	
AlignHorizontal	center
AlignVertical	center
BackColor	white
BackStyle	opaque
BorderColor	black
BorderWidth	0
ClearAfter	No
DataLogging	Standard
DisplayName	
Duration	10000
FontBold	Yes
FontItalic	No
FontName	Courier New
FontSize	18

Name
Uniquely identifies each object.

Summary

4 Samples (1 cycle x 4 samples/cycle)

1 Cycle equals 4 samples

Random Selection

ID	Weight	Procedure	Nested	Stimulus	CorrectAnswer
1	2	TrialProc		X	1
2	2	TrialProc		Y	2

[Stimulus]

Figure 10. Screen shot of E-Studio's user interface.

2. Equipment

Participants viewed the experiment via E-Run on standalone computers in the Human Systems Integration Laboratory at the Naval Postgraduate School. Computers consisted of:

- 24" Dell monitor
- Dell Optiplex 745 desktop computer or Dell (XXX)
  Windows XP Operating System
  Intel Core 2 CPU, 2.66 GHz, 3.0 GB RAM

Experiment slides advanced with the click of any key on the keyboard. Participants wore Altec Lansing Light Studio Stereo (Model AHP524) headphones throughout the experiment.

D. VARIABLES

1. Independent Variable

- Mental model
  - Group A received no specific information about conditions affecting the reliability of the target detection device (TDD).
  - Group B received accurate information about conditions affecting the TDD's reliability; they were told the TDD was 100 percent reliable when no trees or roads were present, 90 percent reliable when only trees or only roads were present, and 80 percent reliable when both trees and roads were present.
  - Group C received inaccurate information about conditions affecting the TDD's reliability; they were told the TDD was 100 percent reliable when no trees or roads were present, 60 percent reliable when only trees or only roads were present, and 20 percent reliable when both trees and roads were present.

2. Dependent Variables

- Trust in the TDD, as measured by responses on Trust Questionnaires
- Mental model accuracy, as measured via responses on Reliability Assessments
- Reliance on TDD, as measured by data logging in E-Prime

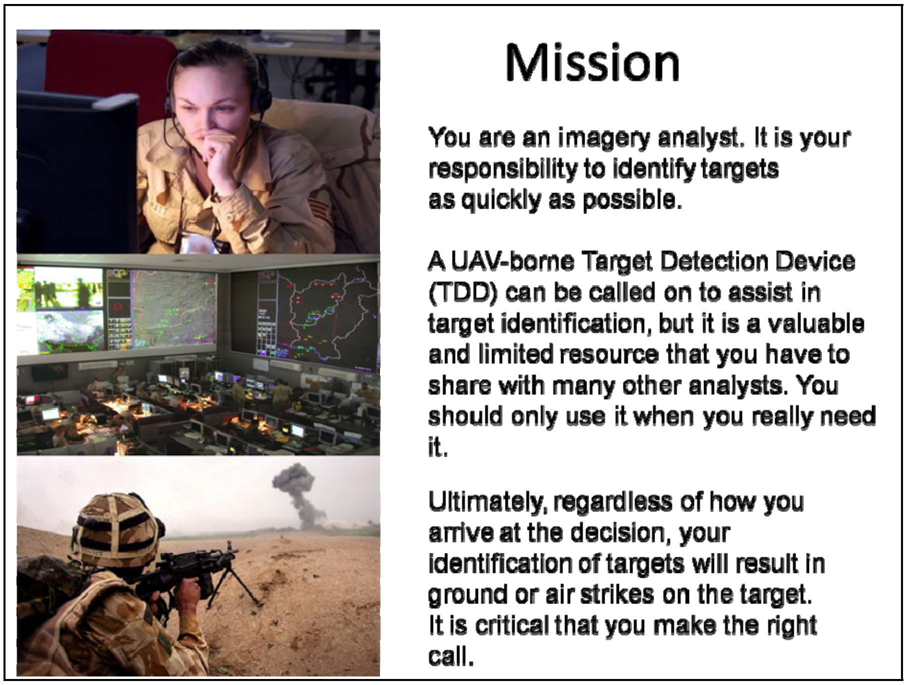
E. PROCEDURE

Participants signed up for a one-hour block of time as their class schedules allowed. Up to three participants could sign up for the same hour. The researcher randomly assigned participants to experimental groups (i.e., the first participant was placed in Group A, the second was placed in Group B, and so on). Participants met the researcher in the Human Systems Integration Laboratory. After signing a form to document their Informed Consent to participate, participants completed a demographic questionnaire (Appendix B).

Next, the researcher presented the participant with one of three reference sheets (Appendix C). The researcher told Group A, “As you go through this experiment, please pay attention to the factors listed on this sheet. After each phase, you will assess whether any of the factors appeared to affect the automated device’s reliability. Please use percentages to indicate how reliable you think the device is under each of the conditions.” The researcher told participants in Groups B and C,

> The sheet in front of you shows the current assessment of the automated device you will be using in this experiment. Its reliability is affected by the factors listed. As you go through this experiment, please pay attention to those factors. After each phase, you will indicate any changes you think are necessary to the original assessment of the automated device’s reliability.

The researcher instructed all participants to wear headphones throughout the experiment. Following these instructions, the participants began the experiment via the E-Run interface. The experiment began with a mission statement, shown in Figure 11, a task description, shown in Figure 12, and example slides, shown in Figures 13 and 14.



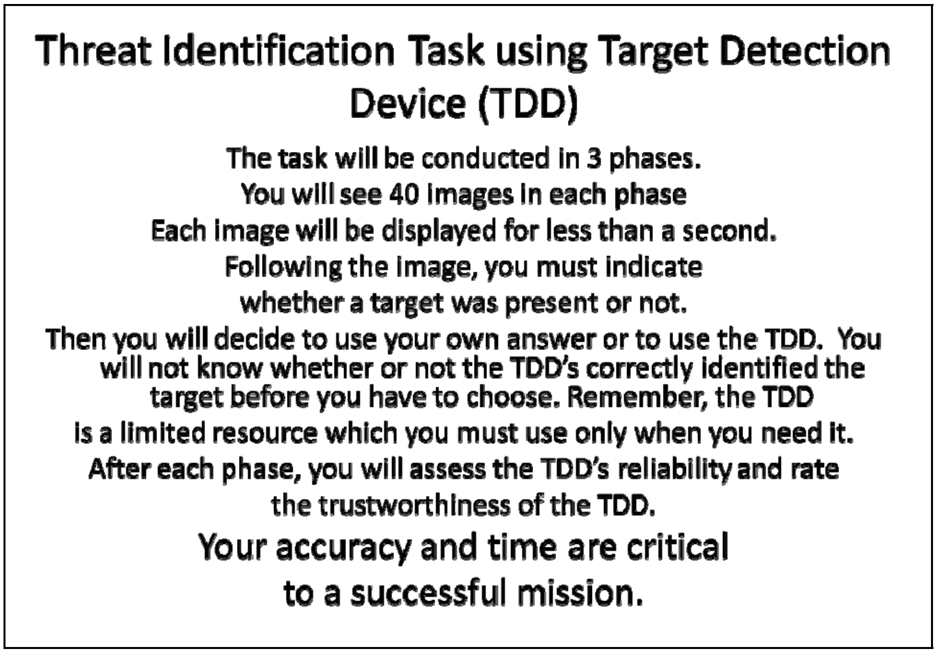
Mission

You are an imagery analyst. It is your responsibility to identify targets as quickly as possible.

A UAV-borne Target Detection Device (TDD) can be called on to assist in target identification, but it is a valuable and limited resource that you have to share with many other analysts. You should only use it when you really need it.

Ultimately, regardless of how you arrive at the decision, your identification of targets will result in ground or air strikes on the target. It is critical that you make the right call.

Figure 11. Mission description slide from experiment.



Threat Identification Task using Target Detection Device (TDD)

The task will be conducted in 3 phases.
You will see 40 images in each phase
Each image will be displayed for less than a second.
Following the image, you must indicate whether a target was present or not.
Then you will decide to use your own answer or to use the TDD. You will not know whether or not the TDD's correctly identified the target before you have to choose. Remember, the TDD is a limited resource which you must use only when you need it.
After each phase, you will assess the TDD's reliability and rate the trustworthiness of the TDD.

Your accuracy and time are critical to a successful mission.

Figure 12. Task description slide from experiment.

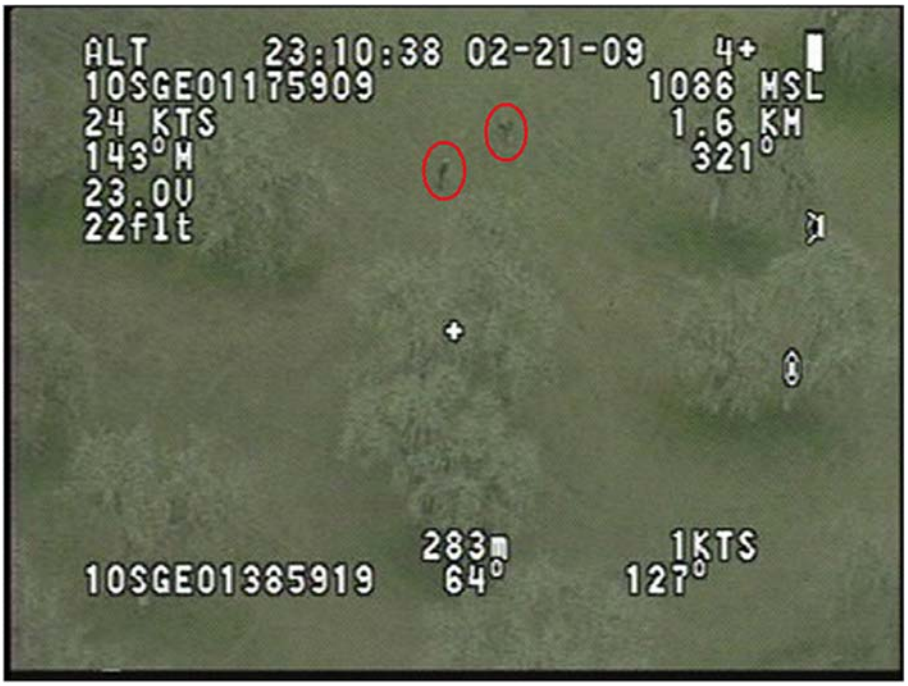


Figure 13. Example of target personnel (circled in red) in still image from experiment.

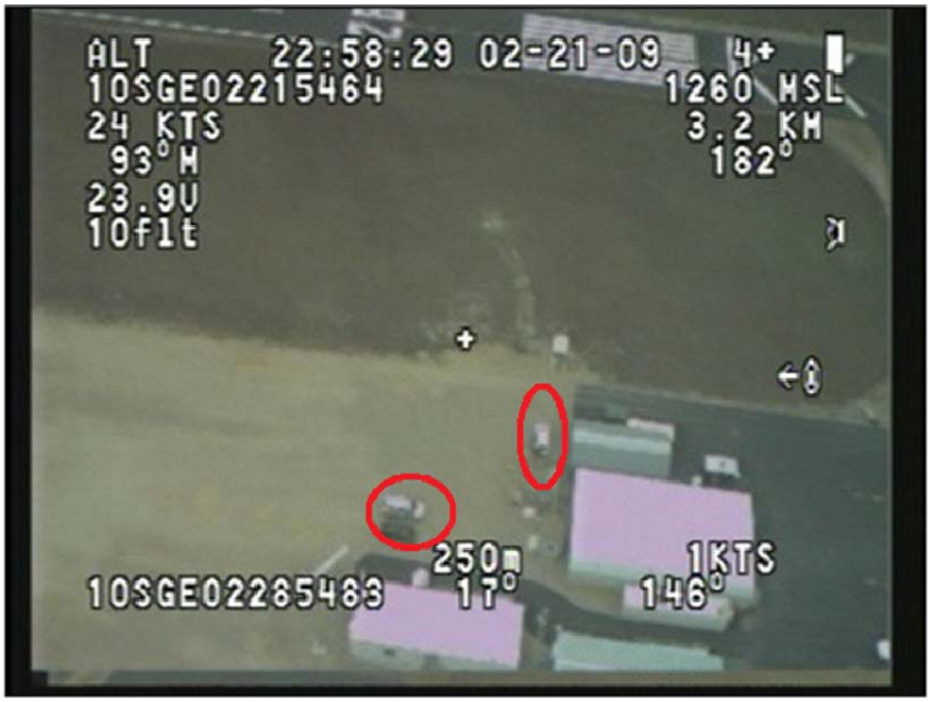


Figure 14. Example of target vehicles (circled in red) in still image from experiment.

After a brief practice run consisting of one image presentation and two decisions, a slide prompted participants to ask questions. The researcher then verified participants' comprehension of task performance for the experiment by asking the following series of questions. The correct answers appear below each question.

- How many images will you see in each phase?
    - 40
- How many phases will there be?
    - 3
- What tasks will you have to do for each image?
    - Indicate target present or absent
    - Choose to use own answer or use TDD
- What tasks will you have to do for each phase?
    - Assess or re-assess TDD reliability
    - Rate TDD's trustworthiness

If participants correctly answered all questions, the researcher instructed them to proceed. If they asked questions, the researcher answered them before participants commenced the experiment. After each of three phases consisting of 40 images, participants rated the TDD's reliability on a blank worksheet (Appendix C) and completed a trust questionnaire (Appendix D). At the conclusion of the experiment, the researcher informed the participants of the TDD's actual reliability, the purpose of the experiment, and the rationale for the intentional deception. Each participant then signed a Debriefing Form (Appendix E) indicating his or her acknowledgment of the intentional deception. The researcher thanked participants for their time and the participants left the laboratory.

IV. RESULTS

A. MENTAL MODEL

1. Scoring of Reliability Worksheets

Each participant completed three reliability worksheets (see Appendix D) from which mental model scores were calculated. TDD reliability varied with each of the four coded conditions. Reliability was set at 100 percent when no roads or trees were present in the image, 90 percent when either trees or roads were present in the image and 80 percent when both trees and roads were present. This reliability stratification remained constant between groups.

The study analyzed mental model accuracy using the mean absolute deviation (MAD) of each participant's answer from true target detection device (TDD) reliability. For example, if a participant rated the TDD's reliability as 90 percent in the first condition, 80 percent in the second, 100 percent in the third, and 85 percent in the fourth condition, the individual absolute deviations from the true reliability would be 0.1, 0.1, 0.1, and 0.05 respectively. The participant's resultant mental model score for that phase would be the MAD, or 0.0875. An accurate mental model should result in MAD scores close to zero.

2. Descriptive Statistics of Mental Model MAD Scores

Table 2 contains the descriptive statistics of mental model MAD scores by phase and group. Figure 15 shows a comparison of the group means of mental model MAD scores by phase. Table 3 depicts the descriptive statistics of mental model MAD scores across phases, and Figure 16 shows the mental model MAD scores for each group across phases.

Table 2. Descriptive Statistics of Mental Model MAD Scores by Phase and Group

	N	Mean	Std. Deviation	Minimum	Maximum
Group A Phase 1 Mental Model	14	.162500	.1992775	.0125	.6500
Group B Phase 1 Mental Model	14	.044464	.0614815	.0000	.2250
Group C Phase 1 Mental Model	14	.146607	.1243738	.0125	.4000
Group A Phase 2 Mental Model	14	.132321	.2084754	.0000	.6500
Group B Phase 2 Mental Model	14	.063929	.0724882	.0000	.2750
Group C Phase 2 Mental Model	14	.124107	.1265732	.0000	.4000
Group A Phase 3 Mental Model	14	.151964	.2029143	.0000	.6500
Group B Phase 3 Mental Model	14	.066607	.0623314	.0000	.2375
Group C Phase 3 Mental Model	14	.111786	.0916065	.0000	.3000

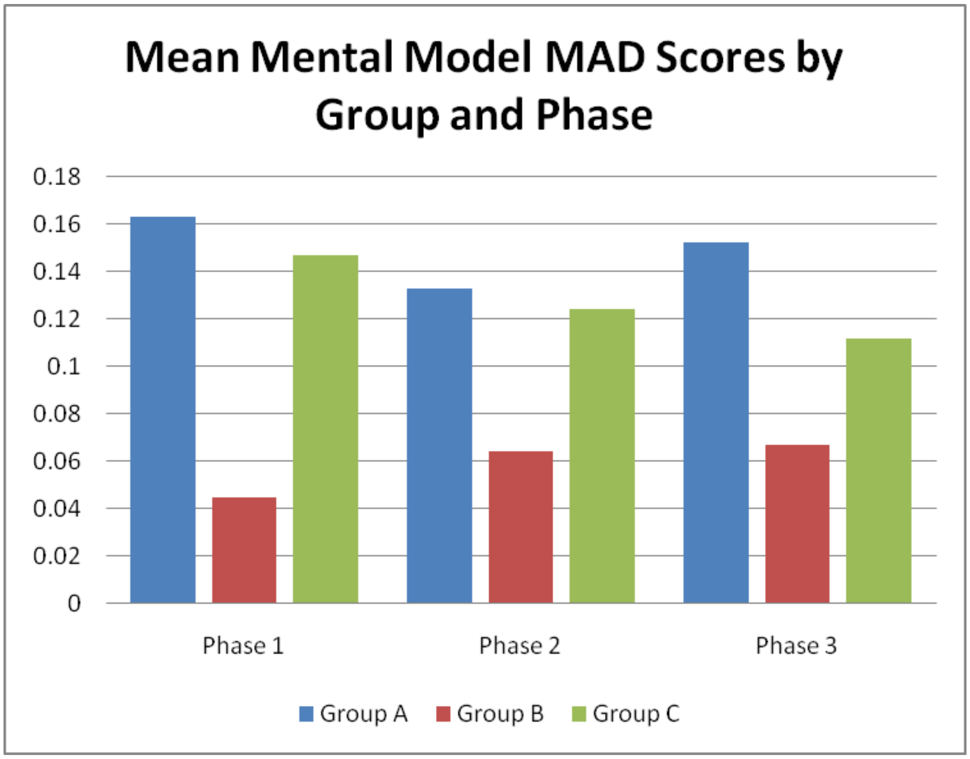


Figure 15. Mean mental model MAD scores by group and phase. Lower values indicate participants' ratings of TDD reliability were closer to the TDD's true reliability. A score of zero indicates perfect agreement between participant's assessment and the truth.

Table 3. Descriptive Statistics of Mental Model MAD Scores by Group

	N	Mean	Std. Deviation	Minimum	Maximum
Group A across phases MAD	14	.148929	.2022524	.0042	.6500
Group B across phases MAD	14	.058333	.0629009	.0000	.2458
Group C across phases MAD	14	.127500	.1024898	.0167	.2917

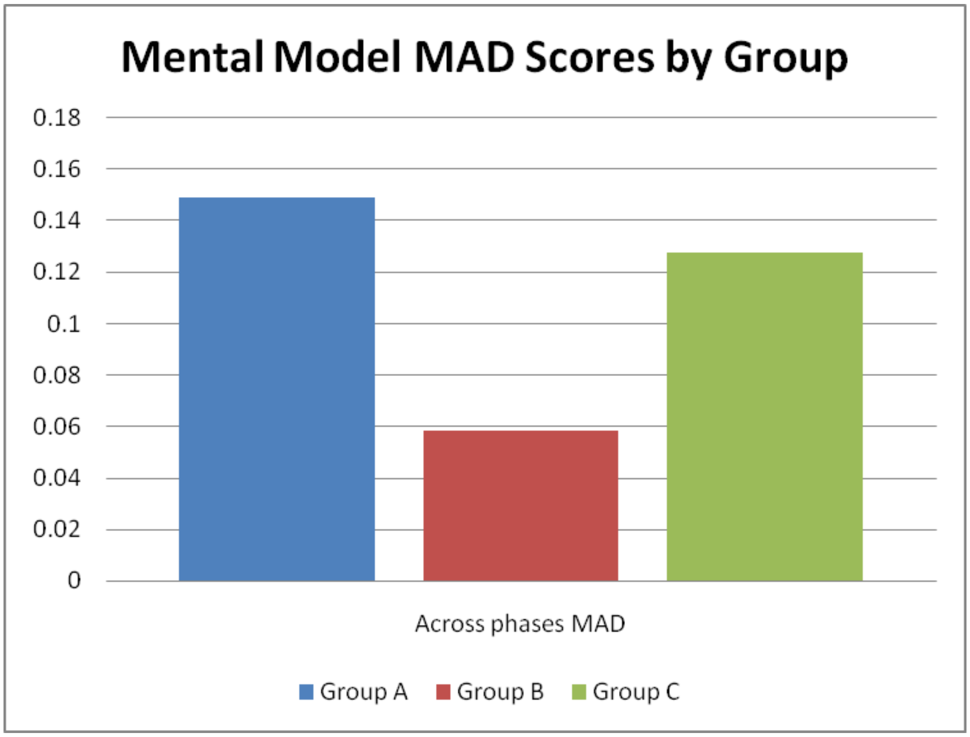
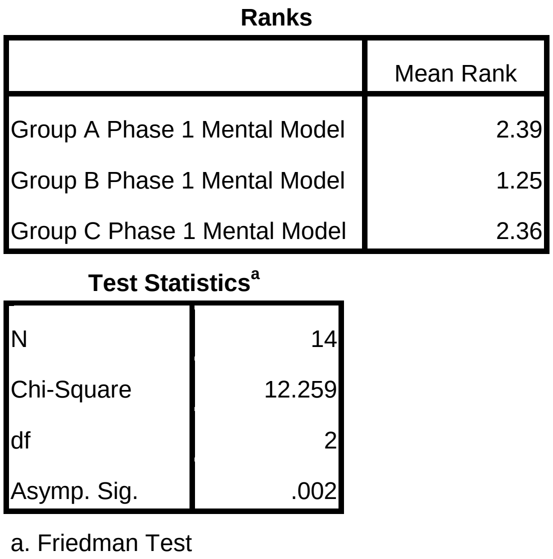


Figure 16. Mental model MAD scores by group. Lower values indicate participants' ratings of TDD reliability were closer to the TDD's true reliability. A score of zero indicates perfect agreement between participant's assessment and the truth.

3. Statistical Analyses of Mental Model MAD Scores

Since the data were not normally distributed and contained quite a bit of variance, mental model scores were analyzed using nonparametric means. Analysis of mental model scores was conducted first within groups by phase, then within phases by groups. Friedman tests indicated no significant differences within groups by phase (Group A p = 0.052, Group B p = 0.074, and Group C p = 0.559). Friedman tests within phase by groups yielded significant differences between groups in Phase 1 (p = 0.002), but not in Phase 2 (p = 0.479) or Phase 3 (p = 0.052). Table 4 shows the results of the Friedman test on mental model MAD scores from Phase 1.

Table 4. Results of Friedman Test on Mental Model MAD Scores after Phase 1



Ranks

	Mean Rank
Group A Phase 1 Mental Model	2.39
Group B Phase 1 Mental Model	1.25
Group C Phase 1 Mental Model	2.36

Test Statistics[a]

N	14
Chi-Square	12.259
df	2
Asymp. Sig.	.002

a. Friedman Test

Finally, because phase contributed little to the outcome, mental model scores were analyzed using each group’s MAD score across all phases. This analysis indicated no significant differences between groups (p = 0.071). Table 5 shows the results of that analysis.

Table 5. Results of Friedman Test on Mental Model MAD Scores Across Phases

Ranks

	Mean Rank
Group A across phases MAD	2.29
Group B across phases MAD	1.50
Group C across phases MAD	2.21

Test Statistics

N	14
Chi-Square	5.286
df	2
Asymp. Sig.	.071

B. TRUST

1. Scoring of Trust Questionnaire

This study used a Trust Questionnaire (Appendix E) developed by Jian, Bisantz, and Drury (2000). Using a seven-point Likert scale, participants rated their level of agreement with a series of statements about the target detection device's (TDD's) trustworthiness. The first five statements are negatively framed, such as "The system is deceptive." The last seven statements are positively framed, such as "The system is dependable." Thus, low agreement levels with the negatively framed questions should indicate greater trust in the system, while low agreement levels with positively framed questions should indicate less trust in the system. Following the Jian et al. example, this study analyzed the results of the questionnaire according to their categorization as responses to negatively or positively framed statements.

2. Descriptive Statistics of Trust Questionnaire Responses

Table 6 shows descriptive statistics of responses to negatively framed questions on the trust questionnaire. Figure 17 shows a comparison of the means by group and phase. On this portion of the questionnaire, lower response scores indicate disagreement with statements such as, "The system is deceptive."

Table 6. Descriptive Statistics of Responses to Negatively Framed Trust Questions within Phases by Group

	N	Mean	Std. Deviation	Minimum	Maximum
Group A Phase 1 Negative Mean	14	2.5429	1.41079	1.00	6.80
Group B Phase 1 Negative Mean	14	2.5571	1.14805	1.00	5.40
Group C Phase 1 Negative Mean	14	2.2714	.85073	1.00	4.00
Group A Phase 2 Negative Mean	14	2.5000	1.34679	1.20	6.80
Group B Phase 2 Negative Mean	14	2.6000	1.15559	1.20	5.40
Group C Phase 2 Negative Mean	14	2.3000	.81430	1.00	4.00
Group A Phase 3 Negative Mean	14	2.5429	1.31366	1.20	6.80
Group B Phase 3 Negative Mean	14	2.3286	1.13573	1.40	5.60
Group C Phase 3 Negative Mean	14	2.2429	1.06750	1.00	4.60

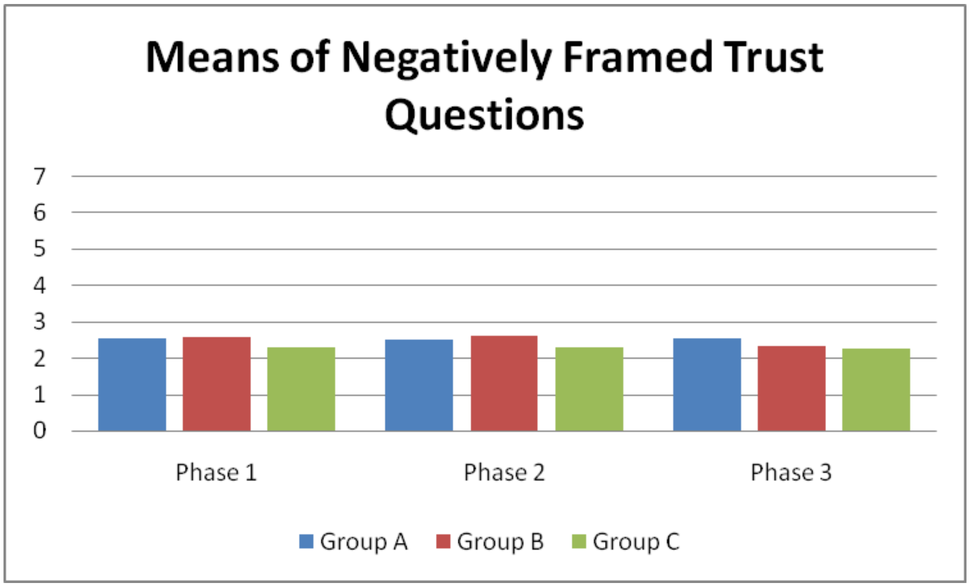


Figure 17. Means of responses to negatively framed trust questions by group and phase. Lower scores on this portion indicate higher trust.

The study next examined the positively framed portion of the questionnaire. In this data set, higher scores indicate greater agreement with statements such as, "The system is trustworthy." Table 7 contains descriptive statistics of those data within phase by group. Figure 18 shows a comparison of the means by group and phase.

Table 7. Descriptive Statistics of Responses to Positively Framed Trust Questions within Phases by Group

	N	Mean	Std. Deviation	Minimum	Maximum
Group A Phase 1 Positive Mean	14	3.9796	1.65261	1.00	6.71
Group B Phase 1 Positive Mean	14	4.2755	1.04147	2.43	5.86
Group C Phase 1 Positive Mean	14	4.7704	1.18142	2.29	7.00
Group A Phase 2 Positive Mean	14	3.8163	1.58726	1.00	6.14
Group B Phase 2 Positive Mean	14	4.5408	1.13206	2.00	6.29
Group C Phase 2 Positive Mean	14	4.6224	1.16005	2.29	6.00
Group A Phase 3 Positive Mean	14	3.8980	1.67672	1.00	6.43
Group B Phase 3 Positive Mean	14	4.6905	1.32543	2.67	6.57
Group C Phase 3 Positive Mean	14	4.6122	1.30140	2.43	6.71

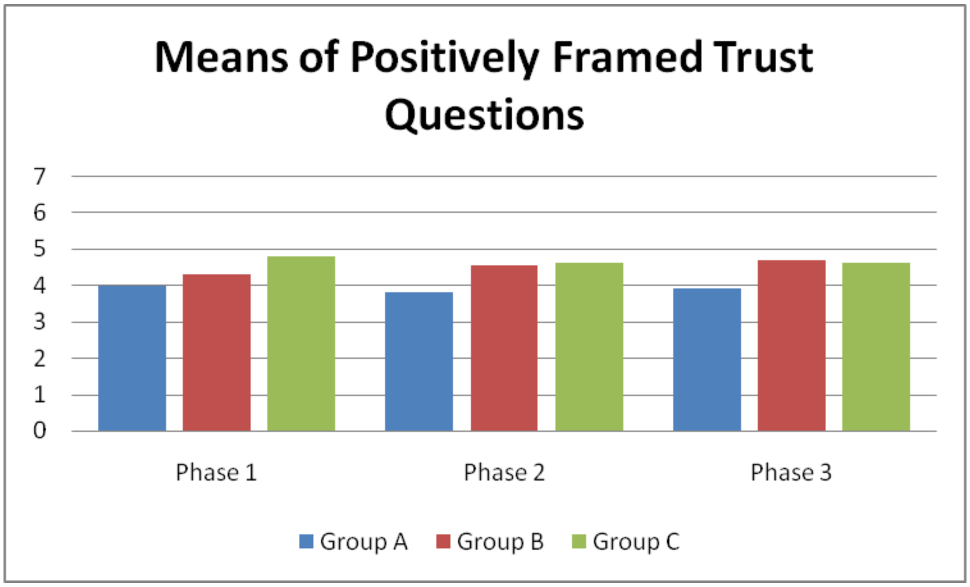


Figure 18. Means of responses to positively framed trust questions by group and phase. Lower scores on this portion indicate lower trust.

3. Statistical Analyses of Trust Questionnaire Responses

Analysis of trust questionnaire data was divided into responses to negatively framed questions and responses to positively questions. Phase had no significant effect on the outcome of the analysis, so data were analyzed only by groups. A Friedman test on responses to negatively framed questions indicated no difference between the groups (p = 0.818). A Friedman test on responses to positively framed questions yielded a significant difference between groups (p = 0.005). Table 8 shows the results of that analysis, and Figure 19 provides a graph of the group means.

Table 8. Results of Friedman Test on Responses to Positively Framed Trust Questions

Ranks

	Mean Rank
Group A Positive Mean	1.60
Group B Positive Mean	2.18
Group C Positive Mean	2.23

Test Statistics[a]

N	42
Chi-Square	10.622
df	2
Asymp. Sig.	.005

a. Friedman Test

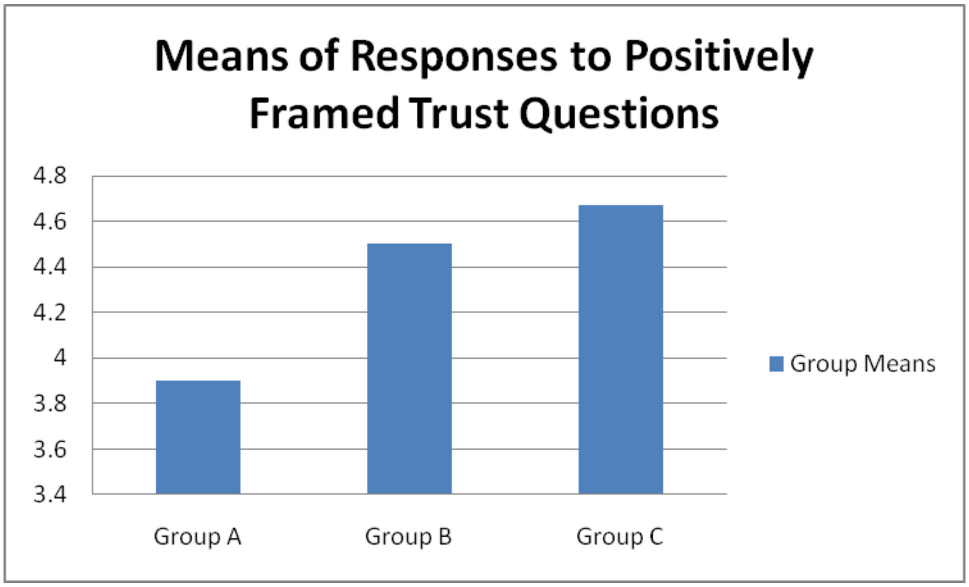


Figure 19. Group means of responses to positively framed trust questions. Y-axis indicates level of agreement from low to high on a seven-point Likert scale.

C. RELIANCE ON AUTOMATION

1. Measures of Reliance

Reliance on the target detection device (TDD) in the study was measured by the participant’s choice to use the TDD or not. After each stimulus slide, the participant pressed ‘1’ to use his or her own answer to the target detection task, or ‘2’ to use the TDD. Each of the 42 participants made this decision 120 times in the experiment.

2. Descriptive Statistics of Reliance Data

Table 9 contains the frequency count and percentages by group and phase for TDD reliance. Figure 20 depicts the total percentage (by group) of trials in which participants used their own answer or used the TDD.

Table 9. Frequency and Percentage of TDD Reliance by Group and Phase

				TDD Use			
				Used own answer		Used TDD	
				Count	Row N %	Count	Row N %
Group	A	Phase	1	362	64.6%	198	35.4%
			2	343	61.3%	217	38.8%
			3	337	60.2%	223	39.8%
	B	Phase	1	398	71.1%	162	28.9%
			2	380	67.9%	180	32.1%
			3	367	65.5%	193	34.5%
	C	Phase	1	355	63.4%	205	36.6%
			2	386	68.9%	174	31.1%
			3	385	68.8%	175	31.3%

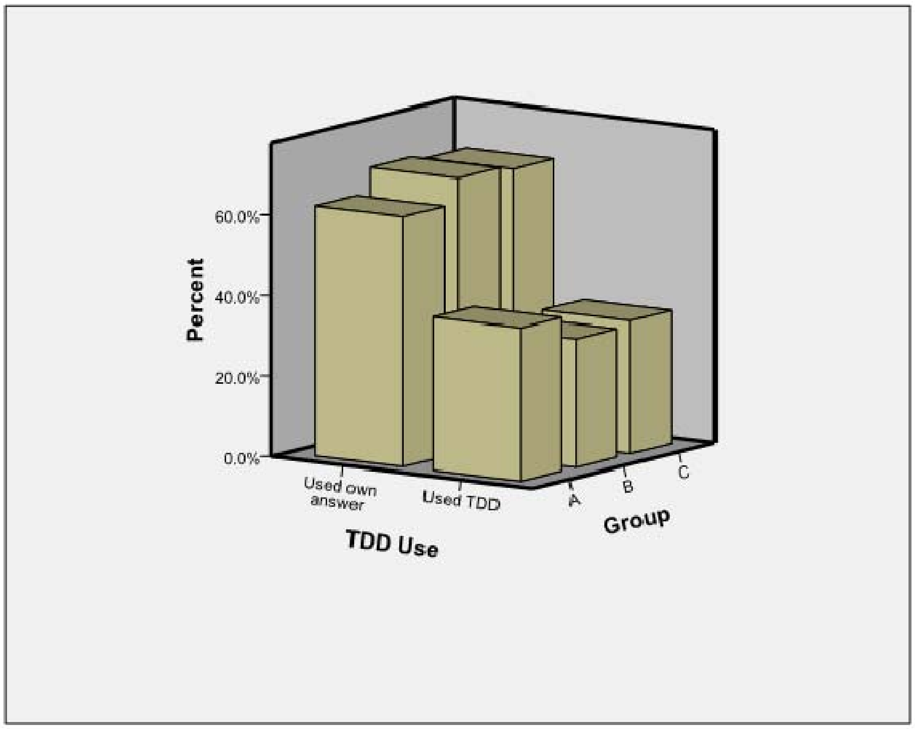


Figure 20. Percent of total trials in which participants used their own answer or used the TDD, by group.

3. Statistical Analyses of Reliance Data

This study used Chi-square tests to examine the effect of group membership on TDD reliance. The analysis indicated a significant difference between the groups' TDD usage (p = 0.002). Table 10 shows the results of that analysis.

Table 10. Results of Chi-square Test on Reliance Data by Group

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15.879	2	.000
Likelihood Ratio	15.775	2	.000
Linear-by-Linear Association	9.321	1	.002
N of Valid Cases	5040		

THIS PAGE INTENTIONALLY LEFT BLANK

V. DISCUSSION

A. HYPOTHESIS ONE

This study found little evidence to support the hypothesis that mental model accuracy varied between groups having little, inaccurate, or accurate information about factors affecting automation’s reliability. Mental model as measured by deviation of participants’ answers from the truth did not vary between groups significantly overall. Although Group A’s mental model appears to have changed over the course of the experiment, there is not enough evidence to support the hypothesis (p = 0.052) nor to determine if their mental models became more accurate or not.

The significant difference found between groups in Phase 1 (p = 0.002) is not surprising since each of the groups had different information at the start of the experiment; at the end of Phase 1 the participants had likely not yet changed their mental models from what they had been told. The fact that no significant differences were found between groups in Phase 2 (p = 0.479) and Phase 3 (p = 0.052) indicates that the between-group mental model differentiation may have been reduced. The reduction between groups could be a result of converging mental model accuracy as participants gained experience with the target detection device (TDD).

Despite the lack of statistically significant differences among the groups, trends in the data bear further examination. Figure 21 depicts the change in group MAD scores over the three phases. As expected, Group B remained the group with the lowest score on mental model deviation. Since this group had accurate information from the beginning of the experiment, it is not surprising that their experience with the TDD did not change their mental model. However, Group A’s mean mental model score vacillated from larger to smaller then back to larger, indicating a possible inability to build an accurate mental model. This may have been a result of that group’s having very little information on which to build a mental model regarding the TDD’s reliability.

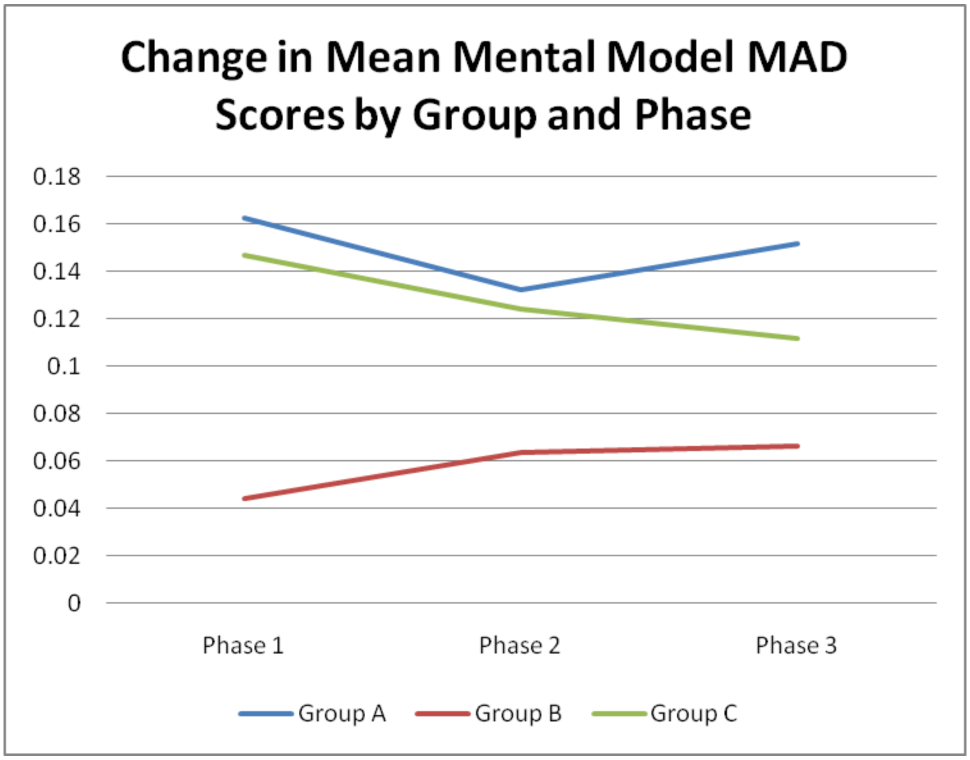
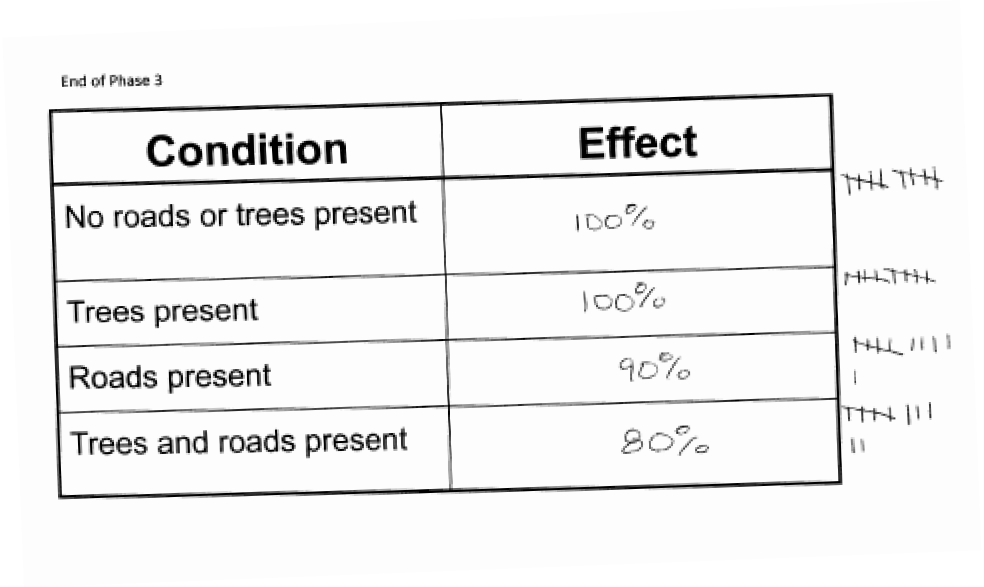


Figure 21. Change in mean mental model MAD scores by group and phase.

Group C's mean mental model score grew increasingly smaller from Phase 1 to Phase 3, indicating the possibility of an improvement in mental model accuracy over time. Participants in Group C may have realized that the TDD was more reliable than they were informed at the start of the experiment. Participants in Group C would have had to factor in their experiences to rate the TDD's reliability and their ratings provide some evidence of a gradual increase in mental model accuracy regarding that reliability.

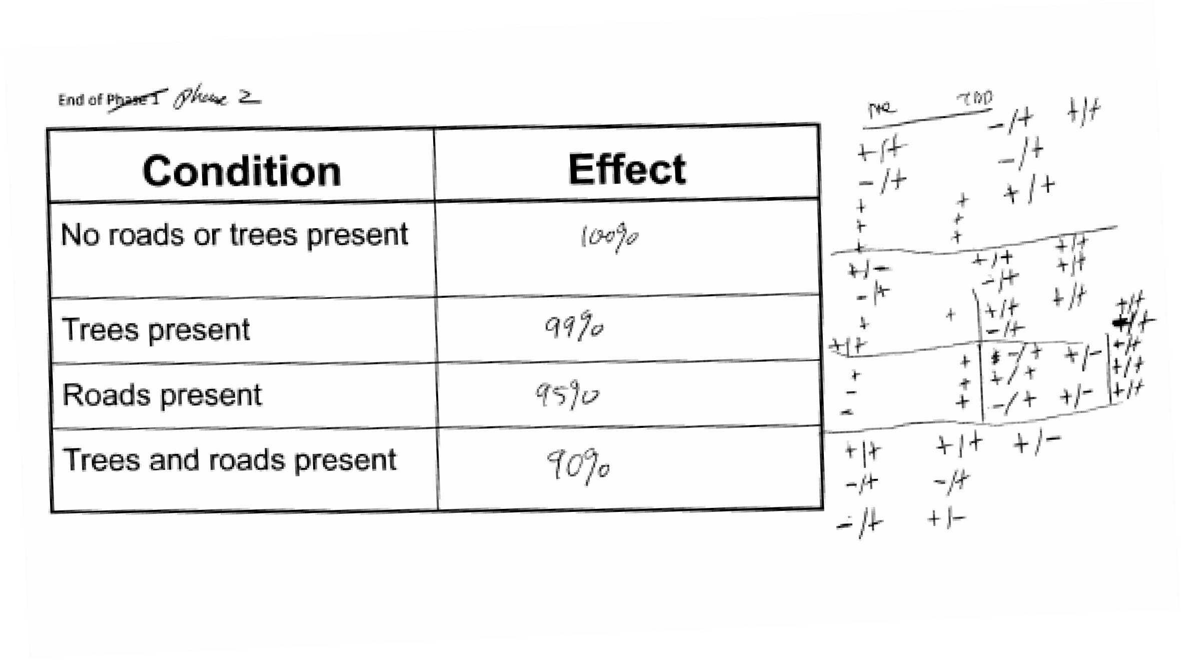
These results support Endsley's (2000) description of mental models as being affected by operators' perceptions, and confirm the complexity of developing mental model accuracy. Post hoc evaluation of participants' reliability assessments showed that many participants kept track of the four environmental conditions and the TDD's record of reliability. Four participants each in Group A, Group C, and five in Group B kept similar notes. Some participants asked the researcher if note taking was permitted, while some did not. The researcher permitted note taking when asked. Representative samples of the notes are shown in Figures 22, 23, and 24.



End of Phase 3

Condition	Effect
No roads or trees present	100%
Trees present	100%
Roads present	90%
Trees and roads present	80%

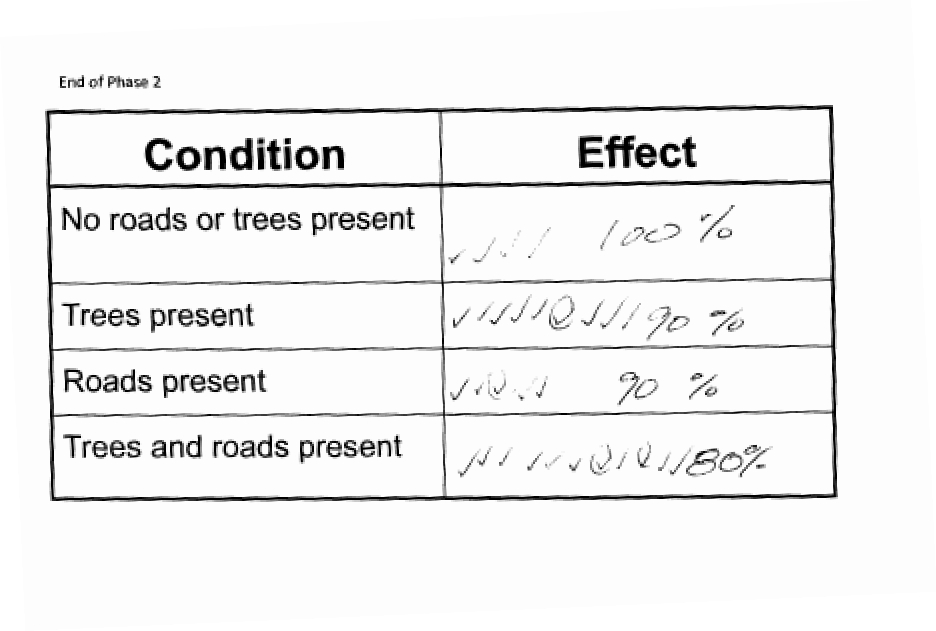
Figure 22. Example of Group A participant's notes on reliability assessment worksheet.



End of ~~Phase 1~~ Phase 2

Condition	Effect
No roads or trees present	100%
Trees present	99%
Roads present	95%
Trees and roads present	90%

Figure 23. Example of Group B participant's notes on reliability assessment worksheet.



End of Phase 2

Condition	Effect
No roads or trees present	✓✓✓✓✓ 100 %
Trees present	✓✓✓✓✓(✓)✓✓✓ 90 %
Roads present	✓✓(✓)✓✓ 90 %
Trees and roads present	✓✓✓ ✓✓✓(✓)✓(✓)✓✓80%

Figure 24. Example of Group C participant's notes on reliability assessment worksheet.

The notes were similar in nature across groups. Table 11 shows the correlation of notes with mental model accuracy by group. Since higher MAD scores indicate less accurate mental models, the negative correlation between MAD score and note taking implies note taking improves mental model accuracy. Thus, the presence of notes positively correlates with participants' mental model accuracy. The notes also suggest that, regardless of the accuracy of information operators have prior to working with automation, they follow similar unwritten rules about how to assess reliability.

Table 11. Correlation between Note Taking and Mean Absolute Deviation (MAD) of Mental Model Scores

	Mental Model MAD	*Notes*
Group A MAD	1	
Notes	-0.42244306	1
Group B MAD	1	
Notes	-0.163960148	1
Group C MAD	1	
Notes	-0.509640181	1

Perhaps the participants who kept notes had higher “resolution,” or greater certainty, because they correctly assessed the specific conditions that degraded the TDD’s performance (Cohen, Parasuraman, & Freeman, 1998). It seems reasonable to suppose that higher resolution regarding the TDD’s reliability is demonstrated in lower mental model MAD scores or higher mental model accuracy.

B. HYPOTHESIS TWO

This study found partial support for the hypothesis that trust varies between groups having little, inaccurate, or accurate mental models about an automated device’s reliability. Although there was no difference between the groups on responses to negatively framed trust statements, there was a significant difference between the groups on responses to positively framed trust statements (p = 0.005). This difference suggests that mental model accuracy affects an operator’s level of trust in an automated system.

The lack of statistically significant differences between group means of responses to negatively framed questions is surprising given the significant difference found in responses to positively framed questions. This result is contrary to Bisantz and Seong’s study, which found no separation between the responses to the two categories of questions (2001). Participant comments may help to explain the current study’s findings. Some participants noted the language in the negatively framed questions was unreasonably harsh. One participant specifically objected to the attribution of emotionally charged words such as “deceptive” and “underhanded” to “a machine.” Similar underlying objections may have caused participants to use the low end of the scale (indicating less agreement with those questions) more frequently. Participants across groups apparently shared this tendency, as the variance in responses to negatively framed questions (1.16) was lower than the variance in responses to positively framed questions (1.71). Additionally, 80 percent of all responses to negatively framed questions occurred in the lowest three scores, as Figure 25 shows.

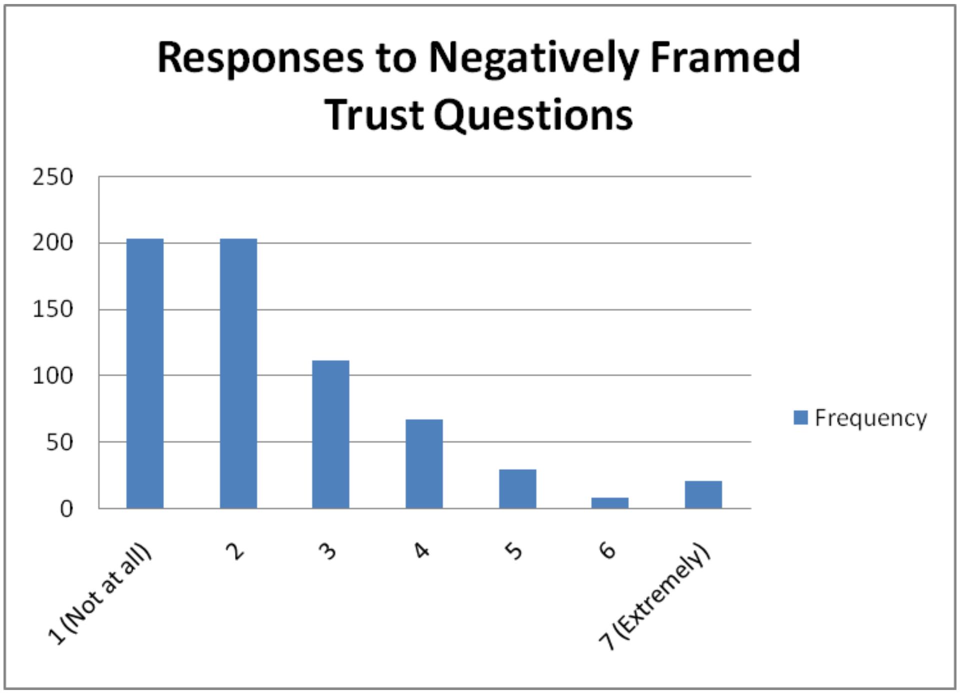


Figure 25. Histogram of responses to negatively framed questions.

Although phase was not a significant factor in the results, the means of responses to positively framed questions over time are worth exploring. Group A, which had little information for an accurate mental model of the TDD's reliability, exhibited the smallest change in trust over the course of the experiment. Group A's lower overall trust ratings could provide support for Dzindolet, Pierce, Beck, and Dawe's work that found operator trust is affected by poor understanding of the automation (2002). Group B, which had the most accurate mental model, showed a trend of increasing trust in the system over time. Group C, which had the least accurate mental model, started off as the most trusting of the TDD but by Phase 3 had noticeably lost some of that trust. Figure 26 depicts the trend of group means over the three phases.

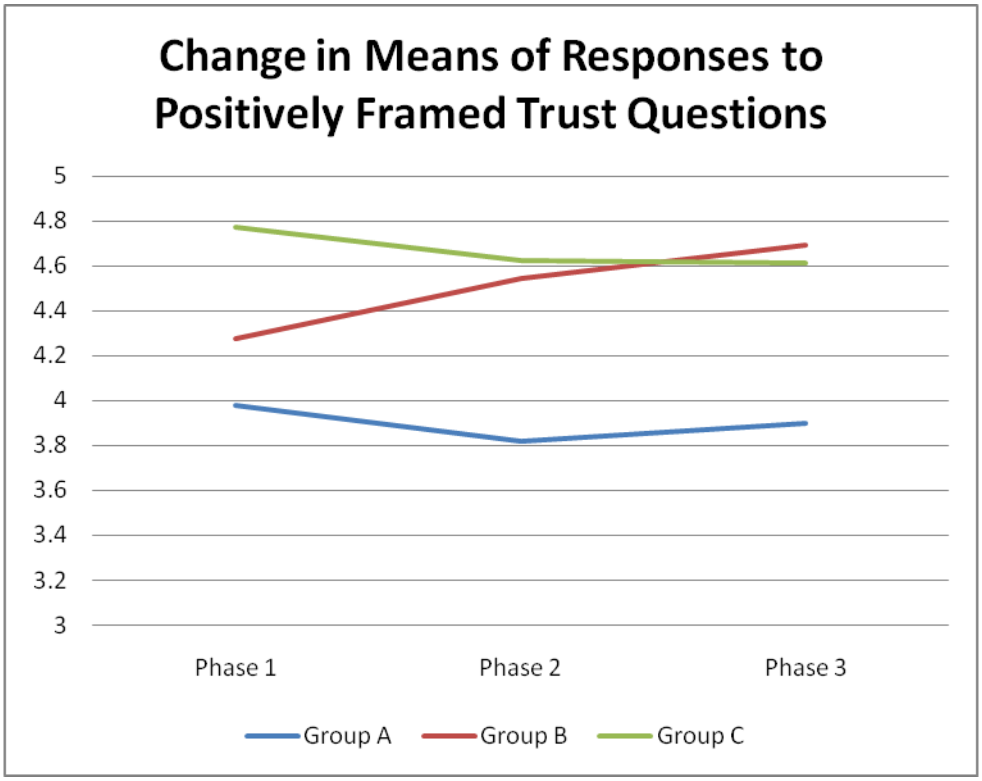


Figure 26. Change in group means of responses to positively framed trust questions. Each phase equals 40 repetitions of a target detection task. Scale on y-axis shows number of points on a seven-point Likert scale indicating agreement with statements about trust in the automated system.

Though not statistically significant, these results are interesting, as they may indicate a measure of trust in the information as well as in the system. For example, if Group C had misleading information about the TDD's reliability, it makes sense that experience with the TDD would show that information to be faulty. Since the TDD was actually more reliable than the participants in Group C were told, we might expect that with experience those participants would develop, or calibrate, their trust in the TDD. However, in this study, participants' trust ratings declined. This result may contradict earlier research suggesting operators calibrate their trust to a level warranted by automation's performance (Lee & Moray, 1992; Muir & Moray, 1996). If that were the case, we would expect Group C to realize the TDD was more capable than first believed and so participants in that group would develop greater trust. On the other hand, this study supports previous findings that indicate a dichotomy between automation performance and operator perception of its trustworthiness (Rovira et al., 2007; Wiegmann et al., 2001). That dichotomy seems evident in the contrasting results from the mental model data and the trust data: although Group C's mental model of the TDD's performance became more accurate, their trust declined.

Additionally, the appearance of declining trust in Group C may indicate the participants were actually rating their trust in the information they received about the system rather than their trust in the TDD itself. This is an aspect of the study that lacks foundation in previous research, but may be an important consideration when measuring trust in similar situations.

C. HYPOTHESIS THREE

Results from this study support the hypothesis that people with little, accurate, or inaccurate mental models about an automated aid's reliability differ in their reliance on the automation. Surprisingly, Group A participants, who had little to inform their mental model of the TDD's reliability, used the TDD more often than their counterparts in Groups B and C. This finding contrasts with an earlier study that suggested operators who have information about why automation might make mistakes increases reliance on the automation, regardless of its performance (Dzindolet, Peterson, Pomranky, Pierce, & Beck, 2003). Conversely, Group A's increased reliance on the TDD may support Nunes' assertion that the need to form a mental model about automation may hinder more than help operators in accomplishing a task (2003). Group A's greater reliance on the TDD may be explained by their ability to focus more on performing the task rather than on assessing the automation. Group A was instructed to assess the TDD's reliability at the end of each phase. Groups B and C were told to compare the given reliability information with their own experience after each phase. Perhaps Groups B and C evaluated their decisions more carefully before using the TDD than did Group A, which would explain some of the difference between groups' reliance.

It is also possible that Groups B and C relied on the TDD more appropriately, although data to support that hypothesis were not analyzed for this study. If the latter were the case, we would expect to see less use of the TDD under conditions known to degrade its performance.

VI. RECOMMENDATIONS AND CONCLUSION

A. RECOMMENDATIONS FOR FOLLOW-ON RESEARCH

Work remains to be done on how to measure the development of mental models and how mental models apply to automation use. The way people think about how automation works appears to influence their use of and trust in an automated system. With the prevalence of automated systems in today's military, a better understanding of the relationship between mental models and automation use will facilitate design of systems and training of personnel.

Additionally, trust is a complex human response to a dynamic relationship between teammates. As military teams increasingly include automated systems, an appreciation for the multifaceted effect of trust on automation use is necessary. Research into the emotional aspects uncovered by this and other studies will provide insights into the human need to trust co-workers (human or machine) at an appropriate level.

Comments from some participants, particularly in Groups B and C, indicated confusion about their role in the experimental task. Some participants wanted clarification regarding how they were to assess the TDD: as an "analyst" or on its performance in the experiment. Taken together, such comments might mean the sample population (Naval Postgraduate School students themselves) is too aware of experimental design and as a result, some participants were trying to provide the right data rather than focus on the experimental task. Experiments to investigate trust in and reliance on automation should be conducted to the maximum allowable extent in operationally realistic environments.

Given the inter-participant variance in this study, pre-testing might help determine people who are naturally more or less inclined to trust automation. This might allow better explanation of variance between participants. Post-experiment debriefing should be conducted using audio or video recording, to capture what participants think about the automation with which they have just worked. In addition, a better experimental design would account for correlations between reliance on automation and difficulty of stimulus.

The stimuli in this experiment were assessed subjectively by the researcher and the timing was tailored as a result of pilot studies. Future experiments should use a difficulty scale for the stimuli in order to analyze response times more accurately.

B. CONCLUSION

This study evaluated three related concepts regarding humans and automation. The results demonstrate that trust, mental models, and reliance are closely related and contribute in sometimes separate but often entwined ways to the performance of a human-machine system. The accuracy of information about an automated system's performance influences its human operator's trust in and reliance on the system. This knowledge alone is enough to warrant continued investigation into the relationship.

This study provides some evidence that between groups with little, accurate, and inaccurate information about an automated device's reliability, there is a difference in trust in and reliance on the automation. Although people with little information about automation's reliability may trust it less, in this experiment, they used it more. This seeming contradiction requires further examination. Additionally, groups with varying levels of information about an automated aid appeared to differ in their mental models initially, but also may have developed more accurate mental models over time.

The results of this study indicate we have much to understand about the interrelationships among mental models, trust, and reliance between humans and automated systems. The complicated nature of each of these interrelated features requires a broader and deeper understanding in order to design, build, and operate effective human-machine systems. In the words of one noted scholar,

> As designers, it is our duty to develop systems and instructional materials that aid users to develop more coherent, useable mental models. As teachers, it is our duty to develop conceptual mental models that will aid the learner to develop adequate and appropriate mental models. And as scientists who are interested in studying people's mental models, we must develop appropriate experimental methods and discard our hopes of finding neat, elegant mental models, instead learn to understand the messy, sloppy, incomplete, and indistinct structures that people actually have. (Norman, 1983)

Although studies involving human participants necessarily mean collecting “messy” data, it is critical that those data are evaluated on their own terms. Only by studying how the nature of human machine interaction evolves can we develop successful criteria for the design of human machine systems.

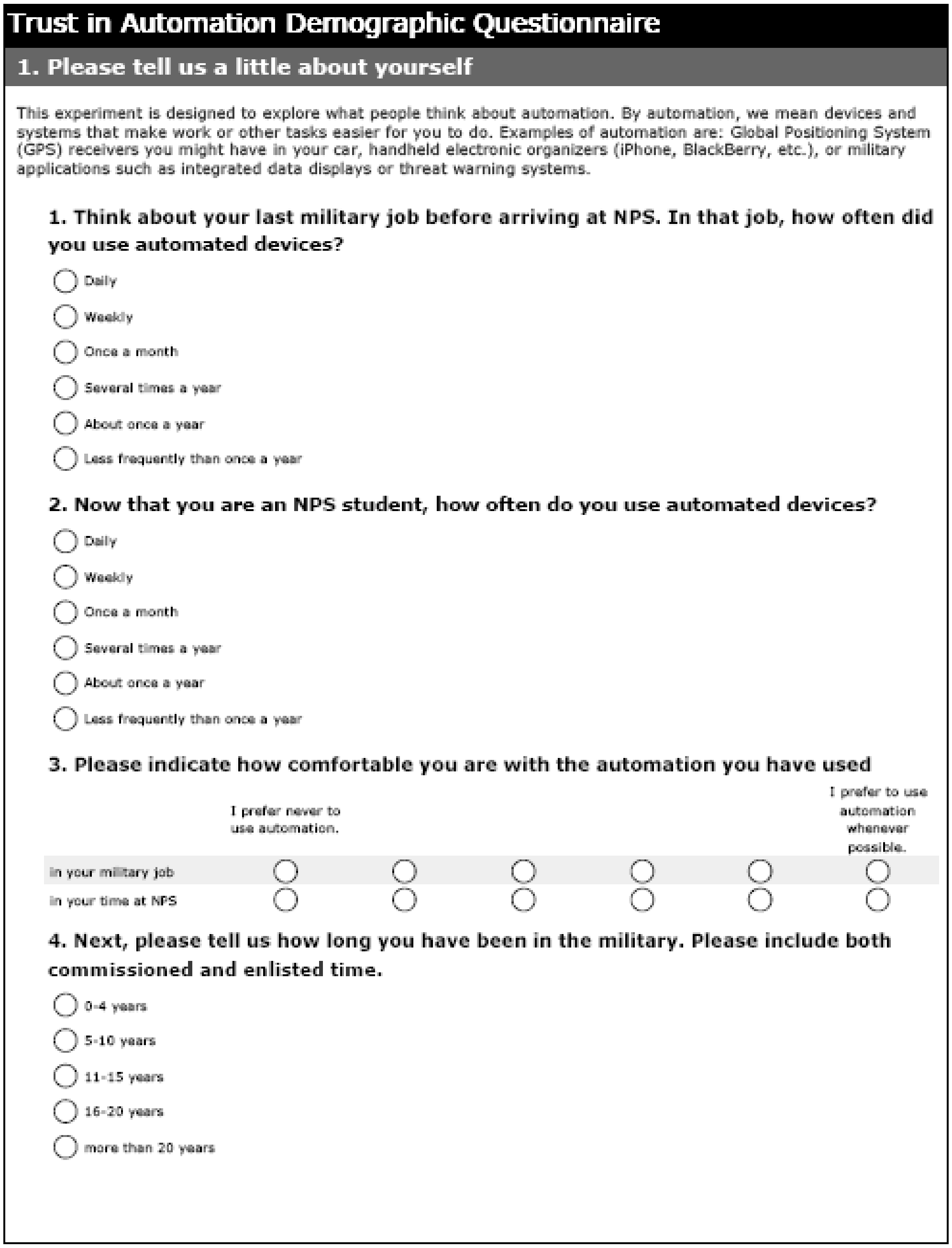
THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX A. STILL IMAGE CODING KEY

Title	Trees	Roads	Threat	TDD Answer	CorrectAns
Still1	y	y	n	absent	n
Still2	n	n	y	present	y
Still3	y	y	y	present	y
Still4	y	y	y	present	y
Still5	n	n	y	present	y
Still6	n	n	y	present	y
Still7	n	n	y	present	y
Still8	n	n	n	absent	n
Still9	n	y	n	absent	n
Still10	y	y	n	absent	n
Still11	y	n	n	present	n
Still12	y	n	n	absent	n
Still13	y	y	n	present	n
Still14	y	y	y	absent	y
Still15	y	y	n	absent	n
Still16	y	y	y	present	y
Still17	y	y	y	present	y
Still18	y	y	y	present	y
Still19	y	n	y	present	y
Still20	y	n	y	present	y
Still21	y	n	y	present	y
Still22	y	n	y	present	y
Still23	n	y	y	absent	y
Still24	n	y	y	present	y
Still25	y	n	y	present	y
Still26	y	n	n	absent	n
Still27	n	y	n	absent	n
Still28	n	y	y	present	y
Still29	n	y	n	absent	n
Still30	y	n	n	absent	n
Still31	y	n	n	absent	n
Still32	n	y	n	absent	n
Still33	n	y	y	present	y
Still34	n	y	y	present	y
Still35	n	n	n	absent	n
Still36	n	n	y	present	y
Still37	n	n	y	present	y
Still38	n	y	y	present	y
Still39	n	n	y	present	y
Still40	n	n	n	absent	n

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX B. DEMOGRAPHIC QUESTIONNAIRE



Trust in Automation Demographic Questionnaire

1. Please tell us a little about yourself

This experiment is designed to explore what people think about automation. By automation, we mean devices and systems that make work or other tasks easier for you to do. Examples of automation are: Global Positioning System (GPS) receivers you might have in your car, handheld electronic organizers (iPhone, BlackBerry, etc.), or military applications such as integrated data displays or threat warning systems.

1. Think about your last military job before arriving at NPS. In that job, how often did you use automated devices?

○ Daily

○ Weekly

○ Once a month

○ Several times a year

○ About once a year

○ Less frequently than once a year

2. Now that you are an NPS student, how often do you use automated devices?

○ Daily

○ Weekly

○ Once a month

○ Several times a year

○ About once a year

○ Less frequently than once a year

3. Please indicate how comfortable you are with the automation you have used

	I prefer never to use automation.					I prefer to use automation whenever possible.
in your military job	○	○	○	○	○	○
in your time at NPS	○	○	○	○	○	○

4. Next, please tell us how long you have been in the military. Please include both commissioned and enlisted time.

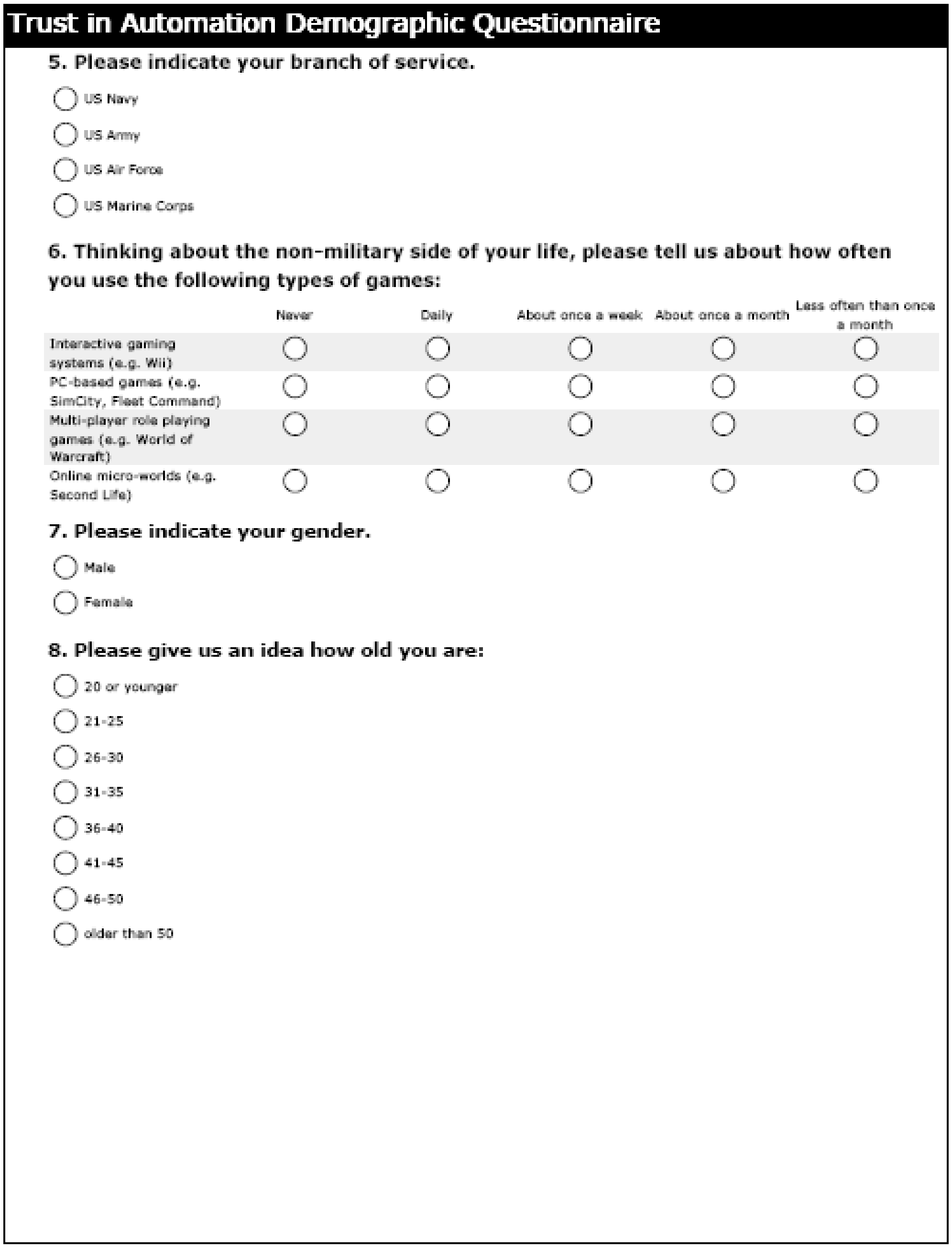
○ 0-4 years

○ 5-10 years

○ 11-15 years

○ 16-20 years

○ more than 20 years



Trust in Automation Demographic Questionnaire

5. Please indicate your branch of service.

○ US Navy

○ US Army

○ US Air Force

○ US Marine Corps

6. Thinking about the non-military side of your life, please tell us about how often you use the following types of games:

	Never	Daily	About once a week	About once a month	Less often than once a month
Interactive gaming systems (e.g. Wii)	○	○	○	○	○
PC-based games (e.g. SimCity, Fleet Command)	○	○	○	○	○
Multi-player role playing games (e.g. World of Warcraft)	○	○	○	○	○
Online micro-worlds (e.g. Second Life)	○	○	○	○	○

7. Please indicate your gender.

○ Male

○ Female

8. Please give us an idea how old you are:

○ 20 or younger

○ 21-25

○ 26-30

○ 31-35

○ 36-40

○ 41-45

○ 46-50

○ older than 50

APPENDIX C. RELIABILITY REFERENCE CARDS

Condition	Effect
No roads or trees present	
Trees present	
Roads present	
Trees and roads present	

Reference card for Group A

Condition	Effect
No roads or trees present	100% reliable
Trees present	90% reliable
Roads present	90% reliable
Trees and roads present	80% reliable

Reference card for Group B

Condition	Effect
No roads or trees present	100% reliable
Trees present	60% reliable
Roads present	60% reliable
Trees and roads present	20% reliable

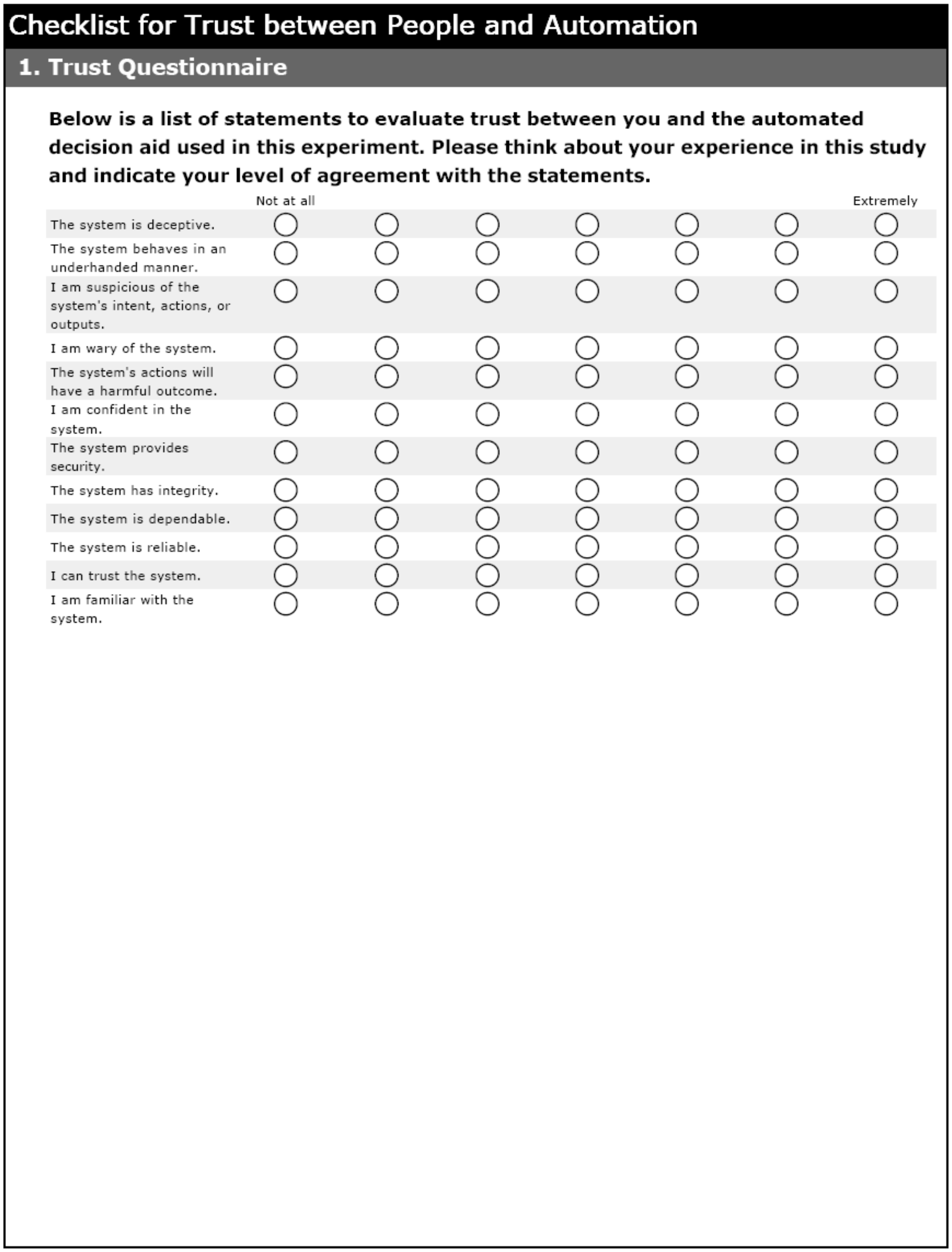
Reference card for Group C

APPENDIX D. RELIABILITY ASSESSMENT WORKSHEET

Condition	Effect
No roads or trees present	
Trees present	
Roads present	
Trees and roads present	

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX E. TRUST IN AUTOMATION QUESTIONNAIRE



Checklist for Trust between People and Automation

1. Trust Questionnaire

Below is a list of statements to evaluate trust between you and the automated decision aid used in this experiment. Please think about your experience in this study and indicate your level of agreement with the statements.

	Not at all						Extremely
The system is deceptive.	○	○	○	○	○	○	○
The system behaves in an underhanded manner.	○	○	○	○	○	○	○
I am suspicious of the system's intent, actions, or outputs.	○	○	○	○	○	○	○
I am wary of the system.	○	○	○	○	○	○	○
The system's actions will have a harmful outcome.	○	○	○	○	○	○	○
I am confident in the system.	○	○	○	○	○	○	○
The system provides security.	○	○	○	○	○	○	○
The system has integrity.	○	○	○	○	○	○	○
The system is dependable.	○	○	○	○	○	○	○
The system is reliable.	○	○	○	○	○	○	○
I can trust the system.	○	○	○	○	○	○	○
I am familiar with the system.	○	○	○	○	○	○	○

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX F. DEBRIEFING FORM

Debriefing Form

During this experiment, an “automated aid” supplied its assessment of the videos under scrutiny and provided its recommended answers to you. All participants were under the assumption that there was an actual automated information aid, when in fact, this is misleading; the researcher programmed responses to the scenarios prior to conducting the experiment and presented the responses as if an automated aid was supplying them. Additionally, some participants received misleading information about the accuracy of automated aid responses.

We did not intend to embarrass you or to insult your intelligence by providing misleading information; rather, we considered the design of this experiment necessary in order to collect valid information about how people construct mental models of automation and how those mental models relate to trust in automation. Prior divulgence of accurate information would have prevented collection of valid data.

If you have any questions or concerns that have not been addressed by the researcher, please contact the Principal Investigator, Dr. Larry Shattuck, 831-656-2473, lgshattu@nps.edu, or the Navy Postgraduate School IRB Chair, LCDR Paul O’Connor , 831-656-3864, peoconno@nps.edu.

Thank you for your participation.

______________________________ ______________
Participant’s Signature Date

______________________________ ______________
Researcher’s Signature Date

THIS PAGE INTENTIONALLY LEFT BLANK

LIST OF REFERENCES

Bisantz, A. M., & Seong, Y. (2001). Assessment of operator trust in and utilization of automated decision-aids under different framing conditions. *International Journal of Industrial Ergonomics, 28*, 85.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (2001). Chapter 12: Shared mental models in expert team decision making. In R. J. Sternberg, & E. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 221-246). Mahwah, NJ: Lawrence Erlbaum Associates.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *Proceedings of the 1998 Command and Control Research and Technology Symposium,*

Congressional Budget Office. (2007). *The all-volunteer military: Issues and performance*. Washington, DC: Congress of the United States.

Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). *A framework of automation use* No. ARL-TR-2412). Aberdeen Proving Ground, MD: Army Research Laboratory.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*, 697.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors, 44*(1), 79.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Presenting misuse and disuse of combat identification systems. *Military Psychology, 13*(3), 147-164.

Endsley, M. R. (2000). Situation models: An avenue to the modeling of mental models. *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society,* Santa Monica. *, 14*(44)

Fahey, L. J. (2007). *Arleigh burke grounding INS plot.* Unpublished manuscript.

Jian, J., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1), 53-71.

Koopman, M. E., & Golding, H. L. W. (1999). *Optimal manning and technological change* (Final No. CRM 99-59). Alexandria, VA: Center for Naval Analyses.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243.

Lewandowksy, S., Mundy, M., & Tan, G. P. A. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied, 6*(2), 104.

Madhavan, P., & Wiegmann, D. A. (2007). Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors, 49*(5), 773-785.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human Factors, 48*(4), 656.

Moore, C. S., Hattiangadi, A. U., Sicilia, G. T., & Gasch, J. L. (2002). *Inside the black box: Assessing the navy's manpower requirements process*. Alexandria, VA: Center for Naval Analysis.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies, 27,* 527.

Muir, B. M. (1994). Trust in automation: Part I. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37,* 1905.

Muir, B. M., & Moray, N. (1996). Trust in automation. part II. experimental studies of trust and human intervention in a process control simulation. *Ergonomics, 39*(3), 429.

National Research Council. (2008). *Manpower and personnel needs for a transformed naval force*. Washington, DC: National Academies Press.

Naval Postgraduate School. *Human systems integration at NPS.* Retrieved February, 2009, from http://www.nps.edu/or/hsi/

Norman, D. A. (1983). Chapter 1: Some observations on mental models. In A. L. Stevens, & D. Gentner (Eds.), *Mental models* (pp. 7-14). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Nunes, A. (2003). The impact of automation use on the mental model: Findings from the air traffic control domain. *Proceedings of the Human Factors and Ergonomics Society 47th Annual Meeting,* Denver, CO. 66-70.

Parasuraman, R., & Mouloua, M. (Eds.). (1996). *Automation and human performance: Theory and applications* (1st ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans, 30*(3), 286-297.

Psychology Software Tools, I. *E-prime 2.* Retrieved March 2009, from http://pstnet.com/products/e-prime/

Riley, V. (1996). Operator reliance on automation: Theory and data. In R. Parasuraman, & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (1st ed., pp. 19). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics, 22*(6), 1296-1308.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of imperfect automation on decision making in a simulated command and control task. *Human Factors, 49*(1), 76.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), *Handbook of human factors & ergonomics* (2nd ed., pp. 1) Wiley. Retrieved March, 2009, from http://csel.eng.ohio-state.edu/woods

Sheridan, T. B. (2002). *Humans and automation: System design and research issues*. Santa Monica: John Wiley & Sons, Inc.

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effect of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science, 2*(4), 352.

Wilkison, B. D. (2008). Effects of mental model quality on collaborative system performance. (Master of Science, Gerogia Institute of Technology).

Wilkison, B. D., Fisk, A. D., & Rogers, W. A. (2007). Effects of mental model quality on collaborative system performance. *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting,* Baltimore, MD*, 31* 1506.

Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors, 31*(6), 617.

INITIAL DISTRIBUTION LIST

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California

3. Lawrence G. Shattuck
   Naval Postgraduate School
   Monterey, California

4. Nita L. Miller
   Naval Postgraduate School
   Monterey, California

5. Patricia S. Hamburger
   Naval Sea Systems Command
   Washington, District of Columbia

6. LTC David Hudak
   Training and Doctrine Command Analysis Center
   Monterey, California

7. Mary T. Dzindolet
   Cameron University
   Lawton, Oklahoma

8. MAJ Bart Wilkison
   United States Military Academy
   West Point, New York

9. Ericka Rovira
   United States Military Academy
   West Point, New York